Review

Perspectives of Machine Learning and Natural Language Processing on Characterizing Positive Energy Districts

Mengjie Han [1,*], Ilkim Canli [2,3], Juveria Shah [1], Xingxing Zhang [1], Ipek Gursel Dino [2,4] and Sinan Kalkan [4,5]

1 School of Information and Engineering, Dalarna University, 791 88 Falun, Sweden; juh@du.se (J.S.); xza@du.se (X.Z.)
2 Department of Architecture, Middle East Technical University, Ankara 06800, Türkiye; ilkim.canli@odtugunam.org (I.C.); ipekg@metu.edu.tr (I.G.D.)
3 Center for Solar Energy Research and Applications (ODTÜ-GÜNAM), Middle East Technical University, Ankara 06800, Türkiye
4 METU Robotics and AI Technologies Application and Research Center (METU-ROMER), Middle East Technical University (METU), Ankara 06800, Türkiye; skalkan@metu.edu.tr
5 Department of Computer Engineering, Middle East Technical University, Ankara 06800, Türkiye
* Correspondence: mea@du.se; Tel.: +46-(0)23-778-628

Abstract: The concept of a Positive Energy District (PED) has become a vital component of the efforts to accelerate the transition to zero carbon emissions and climate-neutral living environments. Research is shifting its focus from energy-efficient single buildings to districts, where the aim is to achieve a positive energy balance across a given time period. Various innovation projects, programs, and activities have produced abundant insights into how to implement and operate PEDs. However, there is still no agreed way of determining what constitutes a PED for the purpose of identifying and evaluating its various elements. This paper thus sets out to create a process for characterizing PEDs. First, nineteen different elements of a PED were identified. Then, two AI techniques, machine learning (ML) and natural language processing (NLP), were introduced and examined to determine their potential for modeling, extracting, and mapping the elements of a PED. Lastly, state-of-the-art research papers were reviewed to identify any contribution they can make to the determination of the effectiveness of the ML and NLP models. The results suggest that both ML and NLP possess significant potential for modeling most of the identified elements in various areas, such as optimization, control, design, and stakeholder mapping. This potential is realized through the utilization of vast amounts of data, enabling these models to generate accurate and useful insights for PED planning and implementation. Several practical strategies have been identified to enhance the characterization of PEDs. These include a clear definition and quantification of the elements, the utilization of urban-scale energy modeling techniques, and the development of user-friendly interfaces capable of presenting model insights in an accessible manner. Thus, developing a holistic approach that integrates existing and novel techniques for PED characterization is essential to achieve sustainable and resilient urban environments.

Keywords: Positive Energy District; machine learning; natural language processing; characterization

1. Introduction

1.1. Background

More and more energy prosumers in single buildings are interacting with the grid. Their frequent energy consumption and injection demands mean that bi-directional grids are increasingly necessary. This scenario is likely to cause a significant redesign of both grid hardware and energy systems [1]. To alleviate this situation, district-level energy management could reduce the volume of interactions with the grid, thus making the energy system more stable. Within the EU, a common agreement has been reached to develop

district-level-based approaches. The Positive Energy District (PED) is one of these methods for improving energy management efficiency, while accelerating the transition to zero emissions in the building sector [2].

Although there have been a number of PED initiatives and pilot projects, the roadmap towards the construction of 100 PEDs by 2025 in Europe is still complex [3,4]. The difficulties comprise an unsettled definition of PEDs [5], the current need for multiple development phases [6], insufficient implementation tools [7], and other technical and non-technical challenges [8]. Many of the projects are still in their early phases [9,10]. Lessons learned from a project on one specific site are not easily transferred to another because of the context discrepancy. Difficulties in understanding the interactions of urban systems for building PEDs are a significant obstacle to applying a single method, making the characterization of PEDs still vague [2]. A technical report of the Joint Research Center (JRC) [11] considers the design a zero-energy district like a PED as a complex process, where multiple stakeholders need to collaborate on a shared target. Ashrafian et al. [8] identified seven factors that are vital to the success of a PED, of which governance is the most significant factor. It needs to be based on a network of interdependency, pluralism, negotiation, and trust for all stakeholders, so that it can then provide support for the other factors. At the same time, the impact of culture cannot be neglected. Different cultural contexts and approaches can mean that cooperation levels between stakeholders can change when collaborating on common goals over a long period [12].

Input data and its customization, grid impact, multi-energy interaction, and information on district infrastructure are the key challenges for energy modeling [13]. Because there is no single tool that integrates all of the necessary information, these challenges exist across the whole of the three-stage PED modeling process: pre-simulation, simulation, and post-simulation [7]. Although external databases can be used, it is not efficient to fit the parameters of these data to a local context. Specifically, research data, operational and observational data, monitoring and evaluation data, and documentation and reusable knowledge are the main types of inputs for modeling PEDs [2]. How to acquire these data and the method to be used for their comprehensive analysis are the fundamental challenges that a PED model thus faces.

1.2. The Importance of PED Characterization and the Need of AI Techniques

Many studies and pilot projects have already sought to develop the idea of the PED and further demonstrate feasible solutions to its existing challenges [14,15]. One area that is particularly important is the activity necessary for scaling up or replicating successful PEDs. PEDs have a number of shared characteristics, even though they are also very specific to their local context. Cities often differ from each other due to various geographic, historical, political, structural, social, legal, and economic factors [16]. Given that the replication of a PED is not simple, it is important to extract the maximum replication potential of a PED when in the early design stages [17,18]. Characteristics of existing PEDs can be used to construct tailor-made solutions for other local contexts. These characterizations can then form an objective foundation from which to construct an efficient PED replication plan. As a result, it is important to characterize PEDs and find the common solutions necessary to boost the replication potential of PEDs for the achievement of climate neutrality and energy surplus.

Existing studies have explored various aspects of PEDs. For example, the densification of residential areas has been recognized as an effective means to reduce both energy consumption and land use for inhabitants within the PED framework [19]. Additionally, management innovation offers a flexible approach in dealing with battery cycling aging, particularly in the context of vehicle-to-building interactions [20]. To facilitate large-scale retrofitting initiatives, providing accurate predictions of potential business scenarios is crucial [21]. Testing facilities play a key role in ensuring the optimal replication of PEDs. They also help identify opportunities for sharing these facilities and enhancing the efficiency of resource utilization [22]. Furthermore, when developing PEDs, it is important

to incorporate citizen engagement methods. These methods help in the understanding of how various efforts can be synergistically combined and improved from a social standpoint [23,24]. Despite these efforts, only a few studies have really looked to characterize what actually comprises a PED [15]. As a guide for the development of PEDs in the future, a deeper understanding of existing PEDs is necessary. This means that more comprehensive scientific knowledge about PEDs and the best methods for running them needs to be gathered. Advanced analytical methods, like artificial intelligence (AI), could be one way to extract the characteristics of PED and, in so doing, generate a virtual PED reference at a more detailed level. This then generates the research question: how can a deeper learning be applied to existing PEDs so that future PED development and implementation can benefit?

1.3. AI Techniques

Energy management strategies, improved energy storage technologies, flexible district energy requirements, and coordinated efforts in cleaner power generation are all required for the energy paradigm shift toward carbon neutrality [25]. It is necessary to demonstrate the efficacy and efficiency of the developed strategies during the design and implementation phases of PEDs through new techniques [26]. PEDs also generate a significant amount of data from various sources: design and construction processes, building services, operational and building management systems, energy infrastructure and transportation systems, and maintenance and replacement systems. The increasing use of digital twins has also facilitated the collection of large amounts of geometric and non-geometric data (building characteristics), weather conditions, and energy data. Analyzing the patterns in these data is crucial to understanding how various systems and infrastructures relate to one another and whether or not they are operating effectively, because PEDs demand the integration of various systems and infrastructures and the interaction between buildings, users, local energy, mobility, and information and communication technology (ICT) systems [27].

AI techniques can be used to analyze large amounts of data and extract valuable insights about the performance of PEDs. These approaches have successfully addressed a wide range of applications, including load predictions, the profiling of energy patterns, mapping regional energy consumption, benchmarking for building stocks, and analyzing the effects of retrofit strategies [28,29]. The ability to learn relationships between input and output makes AI models efficient tools for handling massive and complex data [30]. However, as of yet, there has been no research conducted on how to extract more information about the characteristics of PEDs using AI methods. Even though AI models are being used in more and more areas, the fact that, to date, most studies of PEDs are based on data, models, and model parameters that are unique precludes their conclusions from being broadly applicable. A general framework for AI techniques should be established, so that the performance of PEDs can be more easily analyzed. This should serve to broaden and improve the general applicability of PEDs.

The aim of this paper is to investigate how two AI techniques—machine learning (ML) and natural language processing (NLP)—can be used to analyze the characteristics of PEDs, so that PEDs can be better understood, compared, implemented, and replicated. The rest of the paper is organized as follows. Section 2 introduces the characteristics of PEDs identified in this paper, their nineteen distinct elements. It also outlines the method adopted for the literature review. Section 3 outlines the two AI techniques for analyzing these elements. Section 4 analyzes the reviewed literature based on the elements and the key algorithms. Section 5 discusses the main findings and Section 6 concludes the paper.

2. Elements of PEDs and Research Method

In their PED Reference Framework, the Joint Programming Initiative Urban Europe (JPI UE) defines PEDs as 'energy-efficient and energy-flexible urban areas or groups of connected buildings which produce net zero greenhouse gas emissions and actively manage an annual local or regional surplus production of renewable energy' [31]. This report also

highlights the importance of integrating urban systems with the energy supply and quality of life. However, the process of developing and creating a PED is still an ongoing research topic because of its inherent complexity and uncertainty. For example, diverse climate zones and technology settlements make the optimal location for establishing a PED to be areas in southern Europe that, unlike northern Europe, lack district heating [13]. This section examines the existing literature on PEDs that has in some way sought to analyze the definition of a PED or to identify the key factors or main elements of a PED. The identified elements will then form the basis for a review of the AI techniques proposed to analyze them.

2.1. Elements of a PED

Definitions of PEDs and other related concepts have been shaped by the understanding of sustainable development and the three pillars of sustainability: economic viability, environmental protection, and social equity. Casamassima et al. [32] adopted this tripartite approach to identify six criteria defining a PED: spatial resolution, energy balance, emission, land use, energy efficiency, and energy justice. These were selected because they are independent of geographical location. Land use and energy justice are the main features that distinguish a PED from other similar concepts. PEDs can be seen as contributing to some of the targets identified in the sustainable development goals (SDGs) SDG7, SDG11, and SDG13 [33]. However, the lack of connection of the sustainability dimensions makes the assessment of PEDs fragmented. For a systematic assessment of a PED's contribution to sustainability, the defined Key Performance Indicators (KPIs) need to consider the links between the sustainability pillars [34].

An online-based guideline survey conducted by the IEA EBC Annex 83 group pointed out three clusters of PED elements: the energy system carriers; procedural, institutional, and governance aspects of PEDs; and environmental and spatial quality and the social fabric of PED areas [35]. Bottecchia et al. [36] emphasized thermal and electricity loads in buildings, renewable energy generation, energy management systems, the spatial resolution to identify the boundaries and where to place the generation technologies, temporal resolution, and the objective functions to characterize PEDs, where the multi-objective functions can include energy balance, emissions, cost, and indoor human comfort. Derkenbaeva et al. [37] considered four elements of a PED: geographical boundary, interaction with the energy grid, the energy supply method, and the balancing period. Albert-Seifried et al. [5] compared PED definitions used by European organizations and programs and put forward balanced calculation, boundary, and key energy concepts as the main elements of a PED.

As one of the most important activity and funding leaders, the JPI UE identified three functions of a PED in the urban energy system: energy production, energy efficiency, and energy flexibility [31]. It went on to study these functions in more detail from different angles [5,13,37]. Energy produced for PEDs needs to be sourced from local or regional renewable sources, and achieve the goal of zero emissions. High energy efficiency means the optimal utilization of technologies to reduce energy consumption, which should be prioritized, since the space needed for renewable energy generation is limited in an urban area [37]. Energy flexibility contributes to the resilience and balancing of the regional energy system in consideration of demand management, sector coupling, and storage. It manages the interactions between the systems at different levels. These functions are balanced by consideration of the guiding principles of life quality, inclusiveness, sustainability, and the resilience and security of the energy supply [31].

In summary, by examining the diverse definitions in the existing literature, a common set of PED elements can be drawn up. These elements, rather than offering a narrative description, work to condense definitions by providing a collection of concrete factors that are vital to the success of PEDs and can be individually treated as a means of characterizing PEDs. Nineteen elements have been identified. Each one is either a specification or an enabler of one of the four principles. For example, comfort is an important metric for evaluating life quality and smooth interaction with the grid when the production of

renewable energy can ensure the resilience and security of the energy supply. The elements are grouped into two clusters in Figure 1, with the bottom semicircle being the three functions of the urban energy system and the top semicircle being non-energy issues. Each of the functions forms a sub-cluster, where energy efficiency is on a higher level due to its priority to be considered among the three functions. A similar clustering process has been applied to the non-energy issues. In the upper left quarter, for example, the three pillars of sustainability have been clustered together, alongside a cluster of principle- and policy-related elements. The elements in the top right quarter reflect specific aspects of the local implementation process.
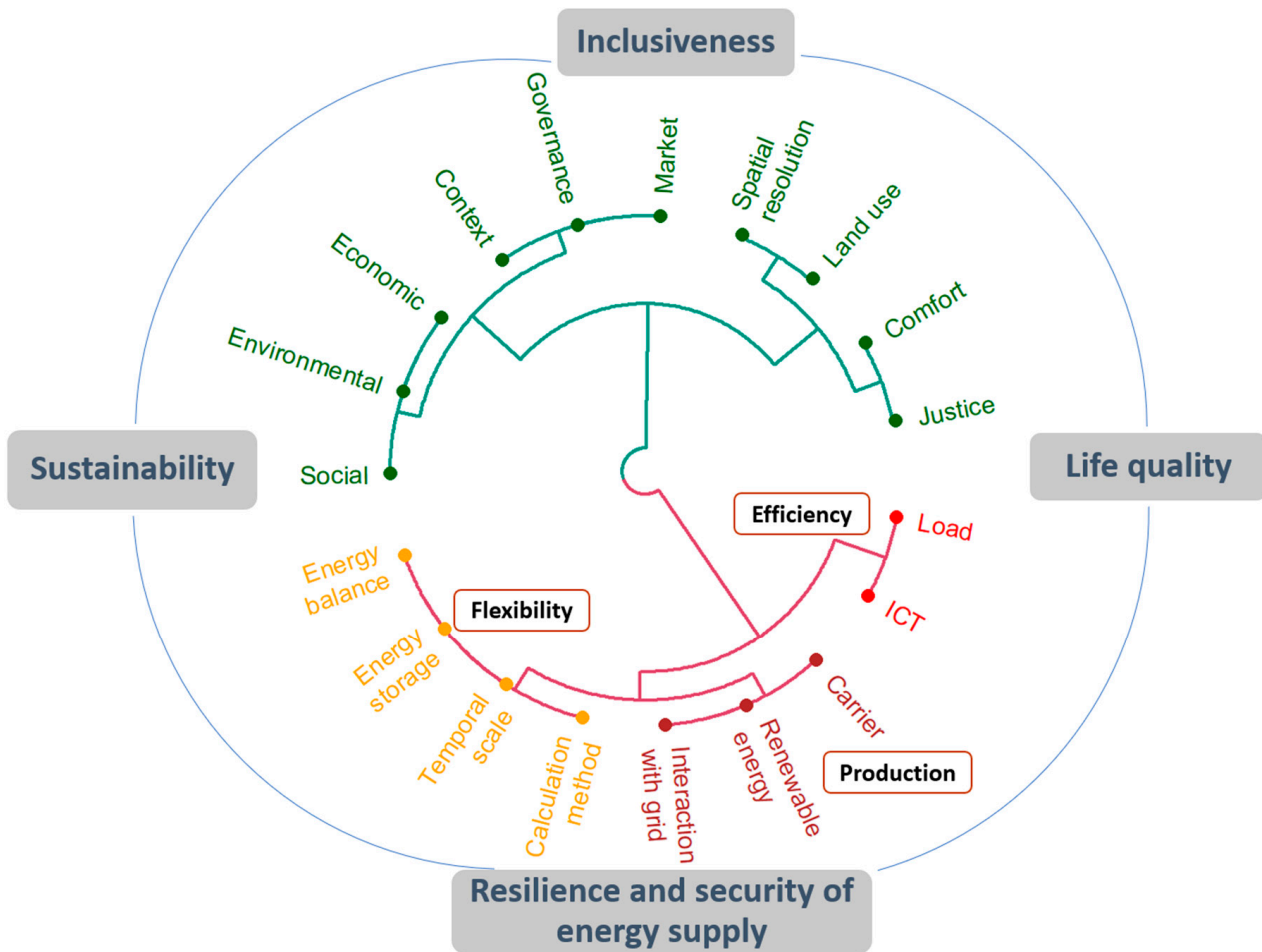

Figure 1. PED elements and guiding principles.

2.2. Research Method

In this paper, we conduct a survey of the literature that has explicitly used ML or NLP to analyze the elements of PEDs. The literature search was carried out in the databases Web of Science, and Google Scholar by combining a PED element and at least one ML algorithm or NLP task in the search term. To be included as an algorithm or task, it was essential that the ML or NLP models could be used for buildings at a district level and had the potential to complement the elements. The large number of papers published on ML has meant that this paper is not an exhaustive search for all of the research related to ML. Only papers published in the last five years (2018–2023) that have had a significant impact are reviewed here. After reading, screening, and elimination of irrelevant papers, we have been able to identify 34 papers related to ML and 37 papers related to NLP. This group of 71 papers forms the basis for the analysis in Section 3.

3. Machine Learning and Natural Language Processing

Artificial intelligence (AI) means designing and applying algorithms in a computational environment to simulate human intelligence and solve complex problems, such as computer vision, control, entity recognition, and classification. As the application of AI in many domains has assisted people and often worked to improve productivity, the integration of AI techniques into building energy management mainly concerns thermal comfort [38] and energy use [39] prediction, building system control [40], fault detection [41], and building information modeling [42]. The evidence from these applications, therefore, provides abundant hands-on experience for PED learning and replication. Among the various AI techniques, the versatility and scalability of machine learning (ML) and natural language processing (NLP) make them highly suitable for large datasets and complex problems. These two techniques are also constantly evolving, with new models and algorithms being developed to improve their performance all the time. Through constructing problem-oriented algorithms, ML learns to automate analytical models from data rather than being taught how to improve its learning abilities. Deep learning, as a subset of ML, is based on multi-layer artificial neural networks that can efficiently model the complex relationships between neurons and recognize complex patterns in the input data. As a sub-field of AI, NLP integrates linguistics and computer science to enable a computer to process and understand natural language data. The most common data source for helping a computer to develop rules for decoding information comes in the form of audio and text. Typical NLP tasks comprise named entity recognition, part-of-speech tagging, topic modeling, machine translation, and text classification. Sometimes, as indicated in Figure 2, NLP tasks need to be executed by using ML or a deep learning method. In these cases, there can be overlaps between solving an NLP task and ML algorithms. Training ML and NLP models requires feeding them with large amounts of labeled or unlabeled data and utilizing an algorithm to optimize and iteratively refine the parameters to improve their performance. The flow of actions necessary for this training to take place—from data collection to final model—are compared in Figure 3 and explained in the following sections.
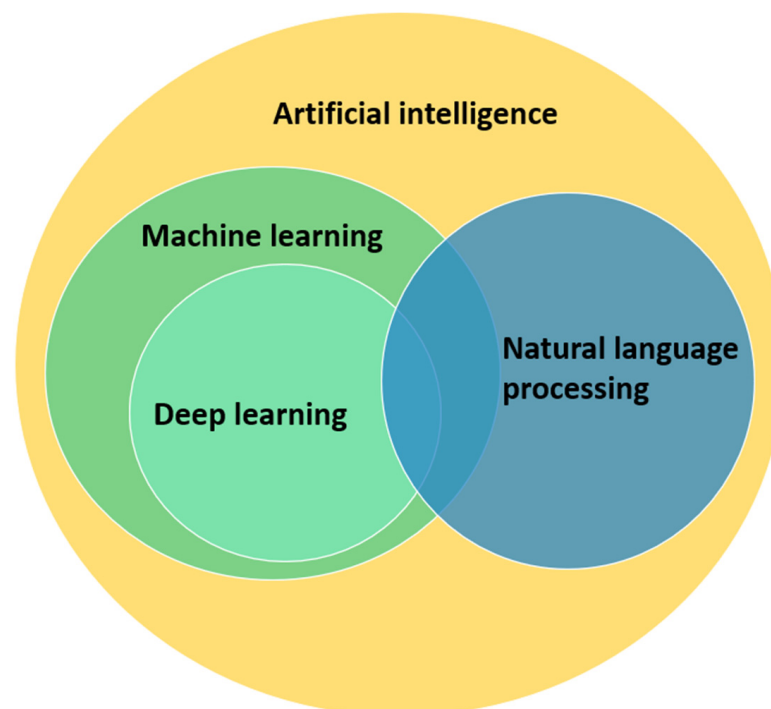

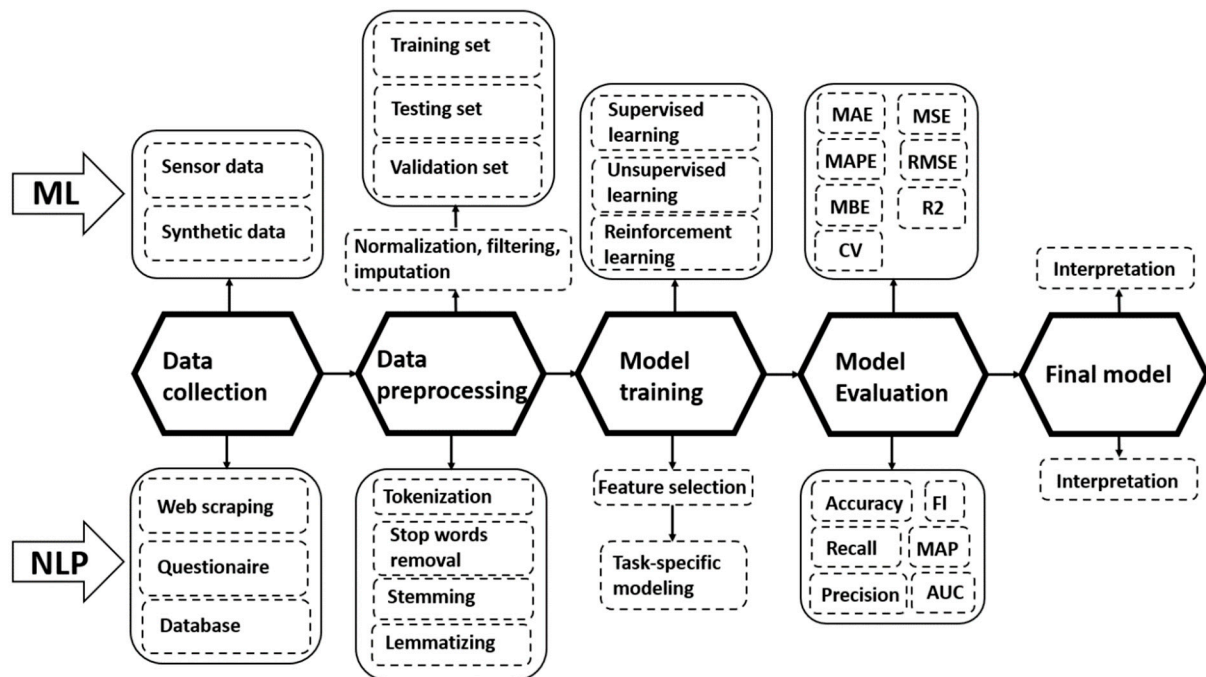**Figure 2.** Two areas of artificial intelligence: ML and NLP.

Figure 3. Flow of actions for ML and NLP model training.

Since the research focus in energy is shifting from single buildings to the district level, a few ML techniques have been developed that seek to understand energy modeling and management at a district level [43,44]. However, studies employing NLP to mine text data are extremely rare [45]. One of the reasons is that the fast growth in big data applications and deep learning architectures occurred at the beginning of the 2010s with the significant increase in computing speed. This breakthrough enables the machine to train millions of parameters to create sophisticated and nonlinear relationships that map the input data. However, it was not until the late 2020s that the execution of NLP tasks by unsupervised pre-trained language models started to outperform human ability [46,47]. By using a huge corpus of input data, these language models are pre-trained to capture the implicit linguistic rules and semantic relationships displayed in human language construction without manually labeling the model output.

3.1. Features and General Machine Learning Processes

3.1.1. Features of ML

PEDs are crucial for a sustainable energy transition to combine high levels of energy efficiency, smart infrastructures, and renewable energy sources in accordance with energy demand [27]. Accurate performance measurement of the different elements that comprise a PED are equally important. Modeling and simulation methods are frequently utilized to assess the current energy use of the building stock, estimate energy demands at various spatial and temporal resolutions, evaluate various design or retrofit options to reduce energy use and environmental impacts, quantify the effects of climate change, and develop energy-efficient strategies for optimization problems [48]. Physical models (also called engineering methods or white-box models) are based on thermodynamic principles and use energy simulation software, for example, EnergyPlus 23.2.0 and IES VE 2023 and detailed input data for energy modeling [27]. Physical models, however, require a high level of operational expertise and extensive modeling data [27,30]. Modeling a PED, therefore, would require a process that can assess the effect of various strategies for evaluating retrofit measures, renewable energy, energy optimization, and implementing effective policies.

ML is a data-driven method designed to complement physical models. Its strength lies in its ability to manage non-linear relationships among data and account for complex interactions and uncertainties. Because they are easy to run, ML models are frequently

employed in energy demand forecasts, the profiling of energy patterns, various retrofit strategies [10], and the forecasting of renewable energy production. For ML, however, gathering the historical data necessary for model training on a large scale is a challenging issue [49]. Data with higher time resolutions, for example, is harder to come by than datasets with lower resolutions, such as quarterly or annual [50]. In cases where no real data exists, where monitoring, for example, has not taken place, ML models can be trained using synthetic data generated from simulations.

Depending on the learning outcome, ML algorithms can be divided into three broad categories: supervised, unsupervised, and reinforcement learning [51,52]. In supervised learning, a model makes predictions for all unseen data points using a set of labeled samples as training data [53]. Applications of this method include using regression to forecast energy usage, predict indoor air quality, compare various retrofit scenarios, control HVAC equipment, and for system management [54]. In addition, classification may handle the energy pattern profiling of particular use cases, the mapping of regional energy consumption, and establishing energy benchmarks for a particular building stock [28]. Unsupervised learning refers to identifying patterns and structure without any prior knowledge and explicit guidance by using unlabeled data [53]. Reinforcement learning is based on a model finding a series of actions to take in a particular circumstance in order to maximize a delayed reward [51,52]. Thus, each of these ML approaches has been shown to be useful for modeling some aspects a PED.

Moreover, the optimization of a PED could be accomplished via a digital twin, a coupled technique for new types of modeling and analysis based on big data and ML/AI [26]. Digital twins based on ML methods continue to evolve in city pilot projects. The potential of a digital PED twin based on ML is to record the dynamic and intricate interactions between the various PED elements and thus to open up new levels of analysis for already complicated energy environments.

3.1.2. General Process of ML

As indicated in Figure 3, the development of an ML model follows the following main stages: data collection, data pre-processing, model training, model evaluation, and final model.

- Data collection—ML requires input data that is mostly collected by sensors or generated synthetically [55]. The Internet of Things (IoT) has a significant potential to enhance the efficiency, effectiveness, and scalability of sensor techniques, while computer simulations and statistical methods are often used for generating synthetic data.
- Data pre-processing—The ranges of the acquired data may be significantly dissimilar from one another or the data may contain outliers, which could lead ML models to inadequate accuracy or learning. In order to tackle these challenges, data pre-processing is used to prepare and transform data into an appropriate form so that useful patterns can be extracted from the data [56]. This pre-processing includes data cleaning, data reduction, data transformation, and data integration [55]. Data splitting can further divide data into training data (for model training), testing data (for testing and evaluating the model) and validation data (for the tuning of model parameters) [57].
- Model training—An algorithm is selected to adjust the parameters in the model based on the training set. The goal is to deploy the trained model on the testing set until it is able to make accurate predictions or optimal decisions.
- Model evaluation—The performance of the ML models is evaluated based on different metrics including, among others, mean absolute error (MAE), mean absolute percentage error (MAPE), mean bias error (MBE), mean square error (MSE), root mean square error (RMSE), coefficient of determination (R2), and coefficient of variation (CV). These performance metrics enable the comparison of different models. Apart from these main steps, hyperparameter tuning sets the optimal parameters necessary to improve the performance of algorithms [58]. Despite being a computationally

expensive process, hyperparameter tuning leads to the increased accuracy, robustness, and reliability of ML models [59].

3.2. Natural Language Processing

3.2.1. Features of NLP

It was not until 2018 that the concept of PEDs started to gain traction. At that time, a number of projects examining alternative ways to carbon neutrality in urban environments published their results. The reports, interviews, publications, and other deliverables issued by these projects provide ample text data that could be interpreted and analyzed by NLP. Until now, it has been the qualitative dimension of these text data that has received attention for PED. Challenges in qualification, data structuring, and semantic extraction have also limited the use of these data. There has been improvement in the use of ML techniques in the field of sustainable development, but there is still considerable potential for the use of NLP in areas like PEDs [60]. For example, one of the current NLP techniques is word embedding, where words are represented as real-valued vectors, so that semantic regularities present in the data can be identified. In addition, NLP is also able to extract knowledge and develop associations between words, which helps in creating clusters of similar words [61]. While topic modeling identifies categories of themes from a collection of documents, pre-trained large language models can perform a wide range of NLP tasks by modifying the structure and finetuning the parameters.

Some pioneering studies have attempted to utilize text data for PEDs. By examining the transcript data and the Knime dashboard in a preliminary analysis of 60 PED projects in Europe, Zhang et al. were able to show the presence of commonly used words and sentiments in the characterization of PEDs [15]. Hedman et al. developed a keyword cloud from research papers on the topic of PEDs or similar [1], which showed that, to date, PED research has focused on building-level innovations, such as zero-energy buildings (ZEBs), intelligent buildings, energy efficiency, and renewable energy sources (RES). Neumann et al. adopted 25 guidelines to shape their collection of opinions regarding spatial scale, audience, and main contents and from this produced a series of keyword and topic clusters [35]. These works have gone some way to setting up the necessary qualitative paradigms for using text mining when characterizing PEDs. However, it is unknown what algorithms were used in these studies and how more advanced algorithms for handling text data might contribute to PED characterization.

3.2.2. General Process of NLP

Apart from the words that carry significant meaning, qualitative data also consists of dates, pronouns, prepositions, and articles that only serve to hamper the working of an NLP algorithm. Data pre-processing, therefore, is a vital early step to perform for any NLP algorithm. It gets rid of unwanted data and identifies the root source, while reducing the size of the data and improving the system's overall performance. Feature extraction and task-specific modeling can then be implemented. The most important techniques in this process are:

- Tokenization—This is a process of separating words, sentences, and phrases into meaningful pieces from a stream of text-making elements called tokens. These tokens are then used as input for further processing. Some textual data contains punctuation marks, dates, and time formats which can create inconsistencies.
- Removing stop words—A number of common words that do not add meaning or generate results in their own right need to be eliminated. Eliminating the most frequently observed stop words such as 'and', 'are', 'the', and 'that' also reduces the size of the data and enhances the performance of the model.
- Stemming—This is a process whereby one word with variant forms is converged into a single 'stem'. For example, the words 'decarbonization', 'decarbonizing', and 'decarbonized' can all be reduced to the single term 'decarboni'. Stemming gives the

root or base word, removes the last few characters of a word, and gives a short word, even though this word usually does not have a meaning.

- Lemmatizing—This is a process similar to stemming but one where the context of the root word is understood and used to generate meaningful representations and aid in information retrieval (IR). Lemmatizing is based on the presumption that an occurrence of the term 'decarbonizing', for example, indicates a connection to data where words like 'decarbonized' and 'decarbonization' are present.
- Feature selection—A process that encodes words into a data type that a computer can use for computation. Bag-of-words, one-hot encoding, and word2vec are all examples of feature selection. Term frequency–inverse document frequency (TF-IDF) is usually used in combination with feature selection to determine the importance of a word in a document.
- Task-specific modeling—In addition to statistical methods, many ML methods can be used for modeling complex problems. A specific NLP task needs to be defined so that the ML model can learn the parameters. When the model is trained, evaluation criteria can be used to measure its performance.

4. Overview of the Literature

4.1. Applications of ML to PED Elements

Applying ML techniques to the characterization of PED elements has mainly focused on energy efficiency, energy production, sustainability/environment issues, and comfort. As can be seen from Table 1, ANN-based approaches, support vector machines (SVMs), and tree-based approaches are considered to be the most important techniques. Other studies involving ML techniques have utilized statistical methods such as multiple linear regression, Bayesian regression, and k-nearest neighbors. In comparison to other PED elements, a large number of applications are found for modeling load and renewable energy.

Table 1. Application of ML to PED elements.

PED Elements	ML Method			
	ANN	**SVM**	**Tree-Based Approaches**	**Other**
Energy efficiency	[62–65]	[63–65]	[62,64,65]	[63–65]
• ICT	[66–70]		[67]	[67,69]
• Load (temporal scale)	[71–77]	[71–75,77]	[73,75–77]	[73,75–78]
Energy production				
• Renewable energy	[79–86]	[79,84–87]	[84–86,88]	[81,84,89]
Sustainability	[90]		[91]	[90,91]
Comfort	[92]	[93]	[93]	[92–95]

4.1.1. Common Approaches in ML

ANN-Based Approaches

Artificial neural networks (ANNs) use artificial neurons to recognize and store information in a manner similar to how brain neurons function [96]. ANNs are a family of methods where the mapping between input and output is modeled as a composition of neurons as simple processing units. A basic ANN structure (Figure 4) involves three main neuron layers: input, hidden and output. Based on the complexity of the modeling task, the number of hidden layers can vary considerably.
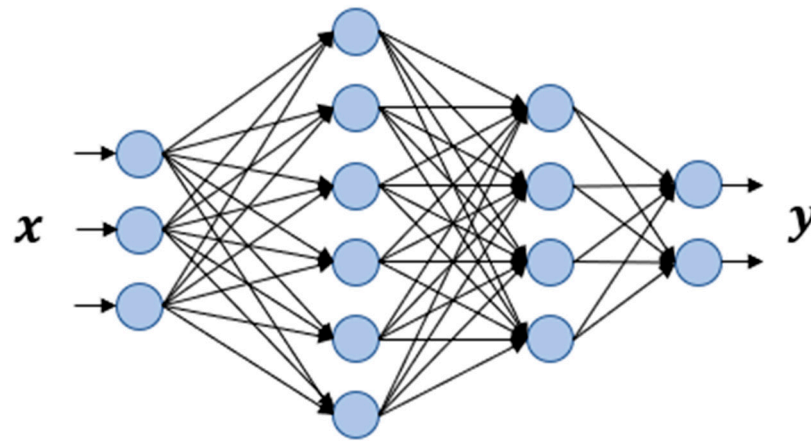
Figure 4. Structure of an ANN where the mapping from input x to output y is obtained through a hierarchical composition of simple processing units, also known as neurons.

In a conventional ANN, the input $x = (x_1, \ldots, x_t)^T$ of a fully connected multi-layer neural network is linearly transformed through multiple layers, with $I_j^{[i]} = g\left(\omega_j^{[i]} I^{[i-1]} + b_j^{[i]}\right)$ for neurons j in layer $[i]$, with $I^{[0]}$ denoting the input x and $g(\cdot)$ a non-linear activation function. The values of each layer are fed forward to the next layer until the output layer turns into y. If $W^{[i]}$ is defined as a weight matrix with the rows being $\left(\omega_j^{[i]}\right)^T$ and $b^{[i]} = \left(b_1^{[i]}, .., b_j^{[i]}\right)^T$, a representation of the layers is:

$$
\begin{aligned}
I^{[1]} &= g\left(W^{[1]} x + b^{[1]}\right) \\
I^{[2]} &= g\left(W^{[2]} I^{[1]} + b^{[2]}\right) \\
&\vdots \\
\hat{y} = I^{[h]} &= g\left(W^{[h]} I^{[h-1]} + b^{[h]}\right),
\end{aligned}
\tag{1}
$$

where $h - 1$ is the number of hidden layers and $\hat{y}$ is the predicted output. A loss function $L(\hat{y}, y)$ is then defined for optimizing the parameters W and b. Due to their ability to learn the complex relationships in a dataset, ANNs have been widely used in forecasting energy demand [55,59,97] and the estimation of renewable energy production [98].

Support Vector Machines

Support vector machines (SVMs) solve classification or regression tasks by transforming data into a high-dimensional space that can be easily separated by a hyperplane [99]. SVMs are designed to maximize the distance between the decision boundary and the nearest data points of each class. This distance is known as the margin, and it plays a crucial role in defining the region of the feature space where the SVM is expected to have good generalization performance.

As shown in Figure 5, the hyperplane is defined as

$$
f(x) = W^T \varphi(x) + b,
\tag{2}
$$

where $f(x)$ denotes the output prediction, W is the weight factor, b is the adjustable factor, and $\varphi(x)$ maps the input space into a high-dimensional feature space. The margin between the support vectors can be easily described as $\frac{2}{||W||}$ and with this, the training objective of an SVM can be formulated as minimizing $\|W\|_2^2$, subject to the samples being classified accurately.
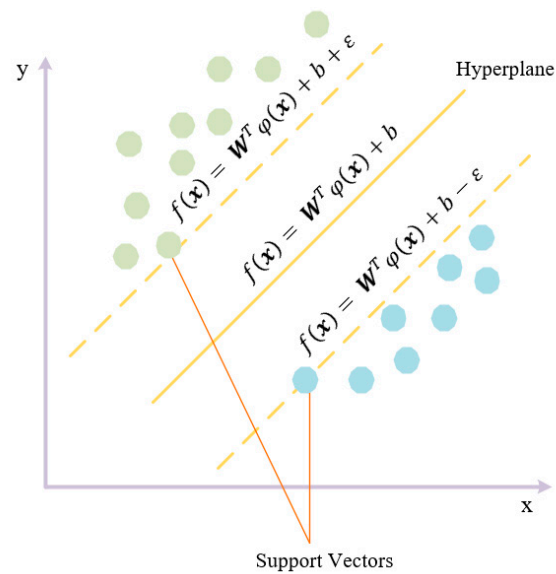
Figure 5. Hyperplane for SVM (dots in colors representing samples of two different classes).

Tree-Based Approaches

Partitioning data into smaller subsets, in a structure similar to the roots of a tree, is another mechanism for solving classification and regression problems. Each non-leaf node corresponds to a single feature, each branch corresponds to a different value for a feature, and each leaf node represents a class of predictions. Random Forest (RF) is one of the most commonly used tree-based approaches. In RF, predictions are made by combining an ensemble of trees in a forest [100]. Given the diversity in any dataset, the combined approach of trees in a forest will lead to more reliable results than a single-tree approach. Figure 6 shows the diagrammatic structure of RF. RF has been used for both classification and regression problems, including modeling solar irradiation, age of building and energy consumption, and forecasting electricity load [101].
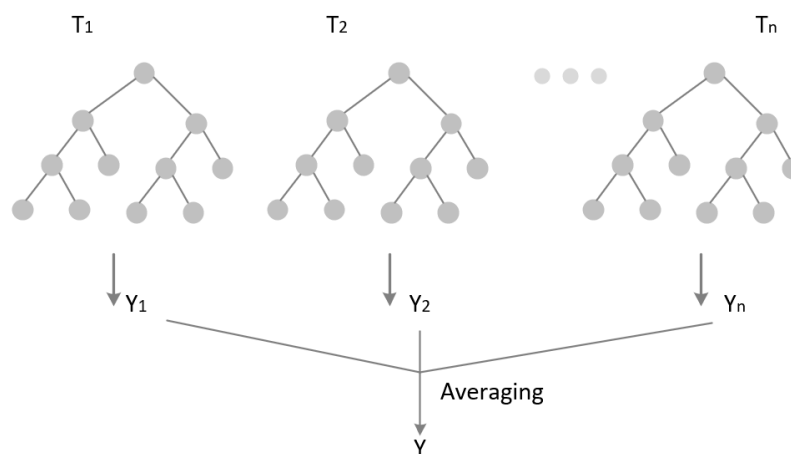

Figure 6. Structure of RF.

4.1.2. Energy Efficiency

The increase in energy consumption and the global energy crisis has increased the importance of energy efficiency research. Efficient energy use reduces energy demand and reduces dependence on external sources of energy for a PED. Studies have explored the use of ML models to improve building energy efficiency over the past few decades [65]. Zekic-Susac et al. attempted to provide an answer as to how to use ML as a key component in effectively managing energy efficiency in the smart city [62]. For this purpose, they utilized deep neural networks, a Rpart regression tree, and RF for predicting the energy

consumption of buildings. Other studies have identified the potential of using industrial data to assist in achieving goals linked to energy efficiency [63]. A framework has been suggested to serve as a reference for process industries on the selection of suitable ML tools for energy efficiency goals. Zhang et al. [64] have examined recent studies that used ML to forecast occupancy behavior and trends with a view to enhancing energy efficiency. They also provided an evaluation of the applicability of different ML algorithms in energy-efficient applications examining occupancy behavior.

Information and Communication Technology (ICT)

'Building ICT' refers to the information technology and communication systems in a building that produce data that can be collected and analyzed so as to improve the building's operational efficiency. ICT is also one of the most important elements of a PED, since it provides stable and reliable connectivity between active and passive devices used for everyday purposes by residents in smart cities [66]. The design, decision-making, and implementation phases of a sustainable and smart buildings system is reliant on an ICT framework, into which an ML method can be implemented [67]. Through the dynamic and ongoing adaptation of network behavior, ML is required to achieve the requisite quality of ICT performance. Data gathered by IoT sensors placed in various locations in a smart city can be utilized by ML to efficiently manage resources and assets [68]. A deep learning model proposed to assess and forecast the performance of an IoT communication system concluded that the model can be useful to modify IoT system behavior [66]. In another study, Serrano integrated reinforcement learning in an intelligent infrastructure model that monitors energy consumption and traffic rates [69]. With the intention of improving energy efficiency, ICT makes it possible to obtain large amounts of data, process these data, and prepare them for application. For example, ML models have been used to classify buildings based on their energy efficiency. These studies have shown how ICT and data science technologies and techniques can be developed for the classification of building energy efficiency [70].

Load

Studies have shown that ML can help evaluate energy load and balances by predicting the heating, cooling, ventilation, and electrical energy demand and load of buildings at both building scale [72,75,76] and district scale [71,74], where 57% of the efforts were carried out at the individual level, and 43% across multiple buildings [77]. Together, these studies contribute to the use of ML-based predictions of energy consumption as a means of evaluating various energy-saving methods. They also provide insight into the use of ML in future energy efficiency applications. Predictions enable demand-side management to perform intelligent control decisions, analyze/balance energy supply and demand, and evaluate the energy flexibility of a building based on smart grid strategies [73].

A number of studies can be clustered around their concern with time granularity. All of them show that it is high temporal resolution data (hourly/sub-hourly as opposed to monthly/yearly) that enables an optimized analysis of the real-time management of energy use in buildings and electrical networks [74,75]. Predictions based on high temporal resolution data can enable utilities to manage resources and also implement strategies to balance the supply and demand of electricity, leading to efficient grid interaction. The studies using high temporal resolution data show that it is possible to evaluate and develop optimization, control, and management for energy-efficient and smart grid applications. For example, minute resolution studies can be used to estimate potential energy flexibility and can enhance both the technical and financial performance of smart grid operations [78].

The research also generally acknowledges that ML methods show higher accuracy performance in the short term and are more effective at making predictions for short time periods than long ones, like, for example, a year or more [74]. While this will be useful in a PED context for short-term energy sharing planning, it is essential for PED development

that accurate long-term predictions are made concerning how to develop an energy supply strategy and capital investment for energy-efficient applications.

4.1.3. Renewable Energy

Renewable energy sources constitute a crucial component of the electricity grid regarding reliability, affordability, and environmental impact [87]. Real-time estimation of produced energy from renewable sources is essential for the planning, management, and operation of electrical power and energy systems. Given the intermittent and unpredictable nature of renewable energy, it is essential to use forecasting as a means of reducing associated risks. This is important for the management and operation of electrical power and energy systems in PEDs. Although applications of ML for renewable energy are mostly related to solar and wind energy predictions [102], studies that benefit from ML models to model energy production from renewable sources can be found across a wider range, including solar energy [79,85,86,89], wind [81,85,86,88], hydropower [82], geothermal energy [84], biomass [83] and wave [80].

4.1.4. Sustainability and Environment

ML approaches for sustainability assessment are essential to help decision makers determine which actions to take to improve sustainability [91], since urban energy systems need to be made more inclusive, safe, resilient, and sustainable [103]. While various methods have been proposed for assessing sustainability performance, fuzzy clustering and supervised ML techniques are more flexible in accepting the large number of sustainability indicators to be used in the assessment of sustainability [90]. Supervised ML techniques have been mostly used for prediction, while unsupervised techniques have been used for the development of new products and materials in the energy sector. However, the availability and refinement of data have been crucial for ML penetration in the energy sector [104]. The main working areas should include engineering, electrical and electronic, computer science, information, and telecommunications. Organizations also need to connect and provide analysis techniques in the data provided by others to create clusters of work and specialization [103].

4.1.5. Comfort

Indoor comfort directly impacts the well-being and satisfaction of the occupants of buildings. If the indoor environment is uncomfortable, occupants may be more likely to make changes to the building's HVAC system or lighting, which could negatively impact the energy efficiency of the buildings. This, in turn, could affect the overall energy balance of the PED. Various ML models can be applied for developing personal comfort systems. Personal comfort models, which are based on the heating and cooling behavior of occupants, can be utilized in daily comfort management practices to enhance occupant satisfaction and optimize energy usage [93]. ML-based control models can also be used to ensure optimal air quality and thermal comfort, while using the least amount of energy from air-conditioners [95]. They manage heat pumps, as well as chilled and domestic hot water storage, for multiple buildings to ensure indoor comfort [94]. This is because an advanced algorithm is able to identify intricate relationships between the air-conditioning systems and thermal environments [92].

4.2. Applications of NLP to PED Elements

The cases where NLP approaches have been applied to PED elements have focused on energy efficiency, energy production, sustainability issues, context, markets, and land use. The techniques used most frequently are word2vec, topic modeling, and BERT and its variations, followed by a number of computationally less intensive techniques, such as POS, text similarity/co-occurrence, and TF-IDF. As can be seen from Table 2, some studies combined multiple techniques to address a specific element [105–113], while some less

frequently used or known techniques were discovered in building ICT, building load, and renewable energy.

Table 2. Application of NLP to PED elements.

PED Elements	NLP Method						
	Word2vec	Topic Modeling	*BERT*	POS	TF-IDF	Similarity/Co-Occurrence	Other
Energy efficiency	[105,114]	[105,106,115]		[106,116]		[117,118]	
• ICT		[107,119]	[108,120]	[108,121]		[107]	[122–124]
• Load	[125–127]	[109]			[109]		[128]
Energy production							
• Renewable energy		[110,115,129]	[130]	[110]	[111]	[111]	[131–133]
Sustainability		[112]	[112,134]				
• Environment	[113]	[113]					
Context		[135,136]					
Market			[137]				
Land use	[138–140]	[141]					

BERT represents the original BERT and its variations.

4.2.1. NLP Techniques

Word2vec

As the most popular implementation of word embedding, word2vec is a technique used in NLP to represent words in a continuous and dense vector space. Two commonly used techniques for word2vec are continuous bag-of-words (CBOW) [142] and skip-gram [143]. CBOW is a model that seeks to capture the semantic and syntactic relationships between words by using the context window of the surrounding words to predict a target word. It learns the representation of words by training on a large corpus of text. The representation of each word is learned by minimizing the cross-entropy loss function [144] upon observing the surrounding context, namely

$$L_{CBOW} = -\log P(w_O | w_{I,1}, \dots, w_{I,C}), \tag{3}$$

where w_O is the output word and $w_{I,1}, \dots, w_{I,C}$ are the context words with length C. Skip-gram, on the other hand, is a generative model that predicts the surrounding context given a target word. Unlike CBOW, skip-gram uses the target word to predict the context words by minimizing

$$L_{Skip-gram} = -\log P(w_{O,1}, \dots, w_{O,C} | w_I). \tag{4}$$

Both CBOW and skip-gram are widely used in NLP tasks such as text classification, language translation, and named entity recognition. CBOW is typically faster to train compared to skip-gram, but skip-gram has been shown to perform better in capturing the semantic and syntactic relationships between words. Both are effective techniques for word embedding. The choice of technique depends on the specific NLP task required and the computational resources available.

LDA Topic Modeling

Latent Dirichlet Allocation (LDA) is an unsupervised method for identifying latent topics in a corpus [145]. It is a way of determining if a group of words within a given dataset semantically relate to each other. LDA uses bag-of-words to consider a document as a vector of word frequency and as a probabilistic generative model for collections of words. In a generative process, a topic is sampled from a topics distribution created for each document. The distribution of words within the sampled topic is then used to select a

word. The process is repeated until the traversal of each word. The observable variables are the words $w_{m,n}$ and the latent variables are the distributions of topics and words.

As indicated in Figure 7, K is the number of topics, M is the number of documents, and N_m represents the length of words in document m. The topics and words are characterized by the multinomial distributions $z_{m,n} \sim Multinomial(\vec{\theta}_m)$ and $w_{m,n} \sim Multinomial(\vec{\varphi}_k)$, respectively. The prior information is given by the Dirichlet distribution $\vec{\theta}_m \sim Dir(\vec{\alpha})$ and $\vec{\varphi}_k \sim Dir(\vec{\beta})$, where $\vec{\alpha}$ and $\vec{\beta}$ are the prior parameters and

$$Dir(\vec{\alpha}) = P(\vec{\theta}_m \vec{\alpha}) = \begin{cases} \dfrac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k - 1}, & \theta_{m,k} \in [0, 1] \\ 0, & elsewhere \end{cases}. \tag{5}$$

The rectangles in Figure 7 indicate replicated sampling from the distributions. Thus, a joint distribution can be formulated [146] as

$$P\left(\vec{\varphi}_k, \vec{\theta}_m, z_{m,n}, w_{m,n}\right)$$
$$= \prod_{k=1}^{K} P(\vec{\varphi}_k|\vec{\beta}) \prod_{m=1}^{M} P(\vec{\theta}_m|\vec{\alpha}) \prod_{n=1}^{N_m} P(z_{m,n} = k|\vec{\theta}_m)P(w_{m,n}|\vec{\varphi}_k, z_{m,n} = k). \tag{6}$$

Various estimation techniques of the parameters in Equation (6) are available in [147].
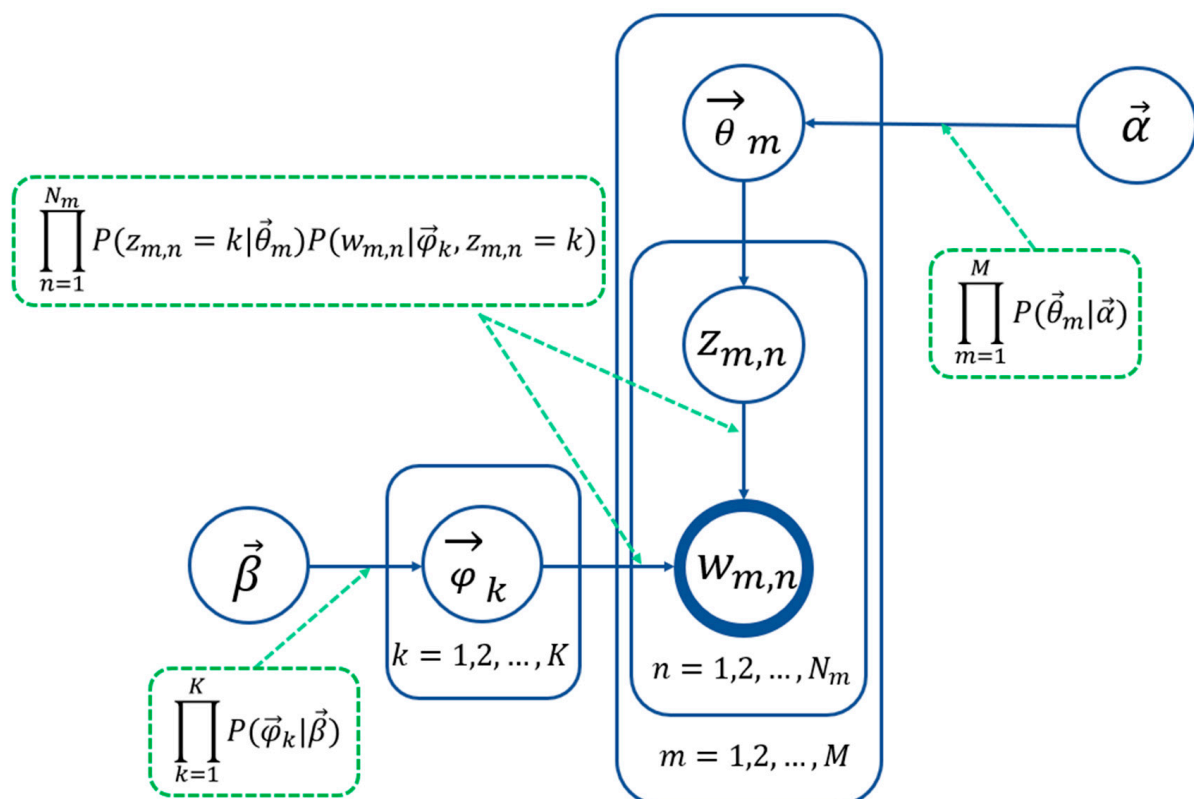

Figure 7. Plate notation of Latent Dirichlet Allocation.

BERT

The Bidirectional Encoder Representations from Transformers (BERT) is designed to pre-train a bidirectional representation of language model conditioning on both the left and right context in all layers [46]. In training BERT, the masked language model (MLM) for predicting words at token level and next sentence prediction (NSP) at sentence level are

used on unlabeled data. In MLM, 15% of the tokens are masked for prediction. To be able to better fine-tune the downstream tasks, the masked token is replaced by a [MASK] token with 80% probability, by a random token with 10% probability, and an unchanged token with 10% probability. For the NSP, two sentences separated by a [SEP] form a training instance, where 50% of the time one sentence is next to the other and 50% of the time the sentences are randomly selected. The trained parameters are then further fine-tuned for downstream NLP tasks with labeled data. For example, a sentiment classification can be indicated by a specific token [CLS]. An illustration of this process is given in Figure 8, where both MLM and NSP are used for pre-training. Multiple downstream tasks are fine-tuned by concatenating a task-specific layer.
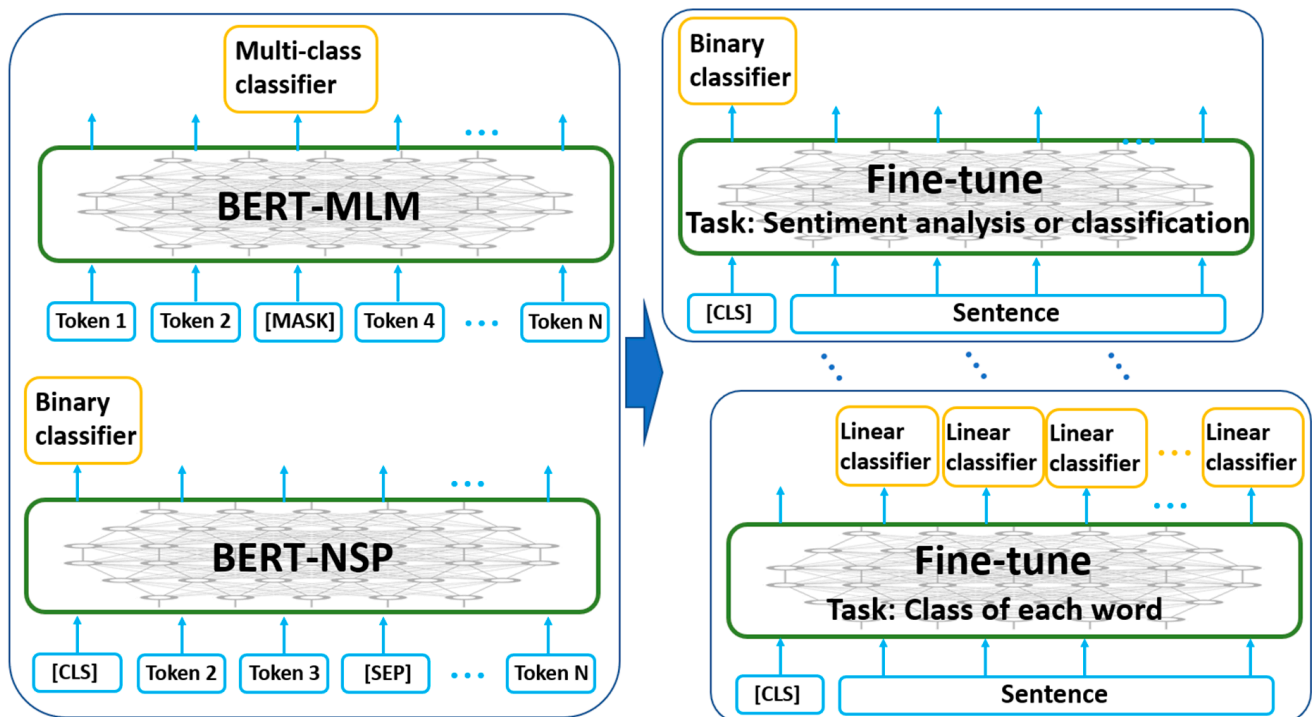
Figure 8. Pre-training and fine-tuning of BERT.

4.2.2. Energy Efficiency

In addition to serving as a corpus for quick co-occurrence analysis [117,118], research papers have also been primarily collected as the document for training word2vec models. For example, Abdelrahman et al. analyzed 30,000 papers with the objective of capturing the relationships between data science and energy efficiency within a network of four categories: data, data science, energy efficiency, and phase [114]. Word2vec was used to represent each word in each category in a vector of 300 dimensions. The consequent usability relation extraction shows that passive design, demand-controlled ventilation, model predictive controls, fault detection and diagnosis, and retrofit analysis use data most frequently, but this is not the case for measurement and verification or operation and maintenance. Moreover, Generative Adversarial Networks, dimensionality reduction, segmentation, and anomaly detection may have potential in modeling energy efficiency. Another consequent analysis of the clustering of concepts based on word2vec identifies key clusters of data science. Similarly, complemented by LDA, word2vec quantifies semantic relationships between keywords for facilitating community detection [105], with heating, ventilation, and air conditioning (HVAC) being the central topics of interest. Factors such as thermal environment, indoor illumination, and occupant behaviors should be considered for modeling energy efficiency.

Energy saving is a key feature of energy efficiency. Wang and Wang identified a topic of household energy-saving technology based on invention patent data and LDA topic

modeling in 31 provinces of China. Sub-topics concern appliances, vehicles, and lighting [115]. One of the challenges of energy efficiency measures is that they are described and organized in different ways [106]. By using 3490 energy efficiency measures from 16 different documents, the topic modeling helps identify six topics, which are distributed in each document, to compare the document [106]. For example, 'ASHRAE Standard 100-2018, Energy Efficiency in Existing Buildings' [148] is 96% identical to 'Energy Efficient Technologies and Measures for Building Renovation: Sourcebook' [149]. 'Commercial Energy Auditing Reference Handbook' [150] is only 14% like 'National Residential Efficiency Measures Database' [151]. In the same study, part-of-speech (POS) tagging was used to reveal the syntactical structure of the documents, where POS learned different parts of speech of each word so that its grammatical role could be extracted. For example, most of the measures are written in the verb–noun format indicating an action–component format. However, description in sentences implies a complex format. In a study of energy conservation measures (ECM), POS tagging was used to pre-process energy audit report data with ECM descriptions. The frequency of each word was then calculated to form ECM dictionaries from the auditors' recommendations. The results were able to guide the implementation of energy efficiency measures by matching the building permit description data [116].

ICT

Among the different sources of data, building type plays a critical role in PED implementation due to its impact on energy performance, regulation, and occupant behavior. Chen et al. applied topic modeling to categorize points of interest (POIs), an approach where building types can be identified based on land use parcels and reclassified POIs [107]. The concept of automated document classification was adopted in [119] for classifying cases of building information modeling (BIM) using topic modeling. The phrases of each BIM use were detected by measuring and comparing the similarities between the definitions of BIM uses and phrases.

Building metadata usually contains physical characteristics, construction materials, systems, and the usage of buildings, which is vital for the maintenance, renovation, and energy efficiency evaluation of a PED. The deployment of smart building applications would be more cost-effective with a standard metadata representation [120]. Waterworth et al. studied a pre-trained RoBERTa [152], an improved BERT, to automatically tag the building sensors with semantic tags based on more than a half million data points on 152 buildings [120]. The authors estimate that, with an accuracy achieving 80%, this approach could assist in the automation of tagging sensor data. In another study, 10 BERT-based and 4 bidirectional language models were compared on POS-tagged building code data to automate regulatory information extraction [108]. A precision of over 95% can be found in a bidirectional LSTM architecture and BERT in combination with error-driven transformational rules. Other applications of NLP to building ICT can also be found for model metadata, such as character-based LSTM as a named entity recognition (NER) model [122] and fuzzy string matching [123].

Load

Modeling building energy load is helpful in improving energy efficiency because it provides information about how energy is being used and where energy savings can be achieved. It is vital for identifying areas where energy is used inefficiently and pointing out how significant energy savings can be achieved, while improving occupant comfort. Based on word2vec, an energy2vec model was proposed by taking time series building load data as its input, with a window length of one minute and the status of appliances [125]. The embedded vectors revealed contextual information about the energy profile by characterizing the habits of the residents and their appliance use. Different lengths of sliding windows of word embedding can be set to decompose the load due to the different operating cycles of the appliances [127]. The issue of encoding categorical attributes and extracting the most relevant ones hinders load prediction. Carrying out word2vec on the attributes before fit-

ting an attention-based LSTM can solve this problem and offers advantages in the medium- and long-term for load forecasting [126]. In addition to word2vec, other techniques have also been used to discover more about trends in energy saving [109] and to outline the ML techniques useful in evaluating energy consumption and intelligent computing [128].

4.2.3. Renewable Energy

PEDs can benefit from renewable energy production as it provides them with access to sustainable and carbon-neutral energy sources. The local production of renewable energy can ensure energy independence and increase the resilience of PEDs. Kumar and Ng conducted topic modeling on 100 research papers to understand the factors of success and growth in renewable energy projects [129]. The identified factors can be prioritized as follows: (1) government policies; (2) public–private partnerships with risk sharing; (3) community support or involvement; (4) positive fiscal mechanisms or terms; and (5) talent. Another analysis on time trends used POS to remove the irrelevant words from 26,533 abstracts and identify more than 1100 topics, with both hot and cold topics in the current research on renewable energy [110]. The hot topics in research may focus on energy storage, photonic materials, nanomaterials, or biofuels, while cold topics may include sustainable development and agriculture. A research task is to find ways to establish and optimize renewable energy systems.

Public acceptance of renewable energy can have a significant impact on the success of an energy transition process. A positive attitude towards renewable energy can lead to increased demand, greater investment, and higher popular support for these technologies. Kim et al. utilized tweet data to fine-tune RoBERTa so as to understand the public stance in the United States towards the adoption of solar energy [130]. The study showed that more positive sentiments were expressed in states with consumer-friendly net metering policies and a more mature solar market. Jeong et al. analyzed more than 18,000 questions registered on the largest portal site in Korea to identify public concerns about renewable energy. TF-IDF was used to place the central words and categories on a word map [111]. A cosine similarity was then conducted based on word2vec to measure how similar two words were. This study confirmed that the public is typically concerned about understanding the characteristics, pros, and cons of different renewable energy resources. Other applications of NLP to renewable energy topics can be found in feature comparisons of two countries using term extraction [131] and techniques categorization of hydrothermal biomass conversion using word-code matrices [132].

4.2.4. Sustainability and Environment

PEDs have been a key strategy for achieving sustainable urban development. In sustainable energy systems, PEDS are seen to reduce carbon emissions and improve energy security and resilience. Saheb et al. conducted a study that combined topic modeling, BERT, and clustering to understand the research focus in AI-based sustainable energy by concatenating the vectors from LDA and BERT [112]. The modeling results revealed that significant attention has been given to sustainable building design and the use of AI to minimize energy usage. On a sentence level, the BERT model showed that it could be fine-tuned for identifying sustainable development goals (SDGs) from multi-source documents [134]. The predicted relational co-occurrence map of the SDGs can support matchmaking, where stakeholders can pinpoint matchmaking candidates and bridge specific needs and proposed solutions. Green buildings are resource-efficient and reduce environmental impact to support sustainability. Current research is focusing on topics such as design, energy saving, rating systems, and life cycle evaluation [113].

4.2.5. Context and Market

Context factors ensure that the design of PEDs is optimized for local resources, policies, and constraints. Topic modeling has been the first choice for capturing these local features [135,136]. In [135], for example, report data on low-carbon transition were used

to understand how local governments interpret the low-carbon transition in China. In general, the three main topics are found to be green industries, decarbonizing the energy sector, and economic growth. By applying a text difference algorithm on the keywords for each pilot, innovative interpretations are outlined to form local contextual conditions, for example, unique resources and infrastructures. In [136], topic modeling was used for disambiguating the objectives of smart city projects between the urban leaders and citizens' needs. By utilizing topic modeling, city officials could effectively engage residents in the development of smart city projects, while establishing a baseline for communication that considers the varying cultural, demographic, geographic, and economic factors within the community. By establishing the necessary infrastructure for managing the supply and demand of renewable energy sources in PEDs, the energy market enables the integration of various energy technologies. By using about 100 thousand website comments data, a concatenation of BERT and bidirectional LSTM could be trained to model public sentiment and thus enable stakeholders to provide accurate technical support in the energy market [137].

4.2.6. Land Use

Efficient planning and land use can promote the development of PEDs. For example, a compact district structure can reduce energy transport and waste. The concept of points of interest (POIs) and the use of GIS data are effective ways to identify land use by analyzing the types and locations of facilities within a particular geographic area. Although quantifying the relationship between spatial distributions of POIs and land use types remains challenging, Yao et al. have developed a shortest path connection to represent sequential POIs for word embedding [138]. While each POI can be seen as a word, each traffic analysis zone is viewed as a document in this study. Zhai et al. have argued that there are limitations to explaining all spatial interaction features of POIs by converting spatial data into sequential document data due to the 2D distribution of POIs [139]. Instead, they considered word frequency to represent POI-type distribution, which shows high performance in clustering functional regions. In another work, cosine distance based on word2vec was used to map POI categories and land use categories, allowing for the quantitative evaluation of the relationship between land use type and its description [140]. As an influential factor, mobility patterns were also used as words in a topic modeling of regional functions to account for the impact of metadata in a region [141].

5. Discussions

The findings identified from this literature review have highlighted the crucial aspects of ML and NLP that can be employed to characterize and reproduce PEDs. As indicated in Figure 9, some key methods, as well as their functions or data requirement for some of the elements, can be identified in this process. According to the number of papers reviewed, it implies that ANN, SVM, and tree-based approaches are three representative ML models, and topic modeling, word embedding, and large language model-based semantic modeling are three primary NLP tasks. For example, the large language mode BERT can be quickly deployed to create a comprehensive picture of a PED and increase semantic interoperability between buildings by analyzing building metadata. This modeling framework can directly enable the effective monitoring and optimization of a buildings' energy use and thus improve energy efficiency. This can be achieved by analyzing building occupancy information, user feedback, and real-time sensor data. Word embedding, through its ability to capture semantic relationships between sequential objects, such as words and geographical positions, can enhance the efficiency of information representation. By mapping out the semantic space of discussions and textual information, stakeholders can gain deeper insights into public opinions, concerns, and expectations regarding PED projects. Another example is the way that deep ANN can facilitate the achievement of indoor comfort through system control in a highly intricate environment. By incorporating an ample supply of data, this sophisticated model structure can effectively capture the nonlinearity of a system, thereby enabling smart control actions to be taken. SVMs, renowned for their

classification capabilities, offer a powerful tool for categorizing energy sources and loads in PEDs. Their effectiveness in high-dimensional spaces makes them indispensable in handling multiple parameters.
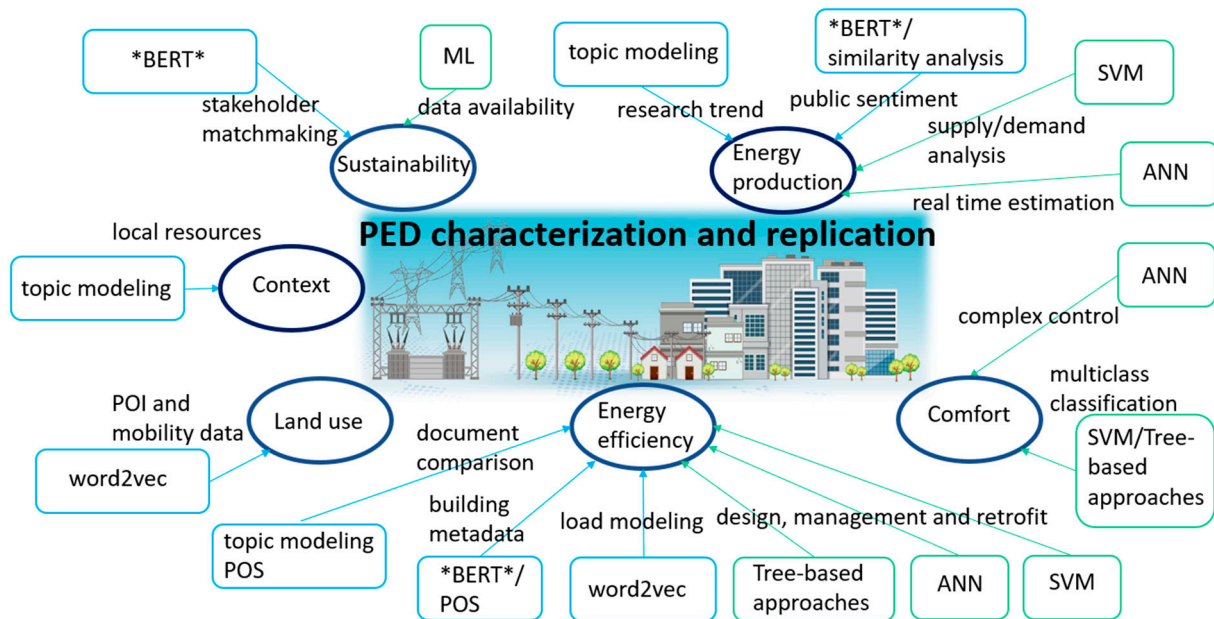

Figure 9. Key features of ML and NLP in PED characterization and replication.

The consideration of these AI paradigms, either jointly or separately, should be dependent on the phase that a particular PED is currently in. For example, during the design phase, stakeholder matching could be used to identify the most effective sustainable design strategies, while evolutionary algorithms could be used to optimize the placement and sizing of renewable energy systems in the implementation phase.

Several gaps can be found in the literature seeking to characterize PEDs. Firstly, there are still a number of elements that have not been considered from the perspective of ML or NLP, such as renewable energy carriers, integration with the grid, energy balance, or justice. One of the reasons for this is that there has been an insufficient amount of data to quantify them. Decision making is thus based on qualitative information. Another reason is the absence of a shared standard for these elements and a limited practical experience in defining them. Thus, additional efforts are required to develop unified methods for defining and quantifying these elements in a PED ecosystem, as well as for disambiguating the concepts in PED replication. Consequently, to facilitate the characterization of PED elements, it is essential to not only invest in collecting higher-quality data but also to focus on efficiently pre-processing these data, which is still a limitation of AI methods.

Secondly, since PEDs involve various spatial and temporal resolutions, as well as the relationships between buildings/urban infrastructure and energy systems at a district level, urban-scale modeling approaches should be considered within the scope of PED research. One of the goals of urban energy modeling is to optimize energy systems, while promoting sustainability and resilience. The modeling experience of ML-based predictive analytics and energy system optimization and NLP-based regulatory analysis and sentiment analysis can be seamlessly utilized for designing energy-efficient PEDs, developing local renewable energy strategies, and evaluating the impact of PED development on carbon emissions [153].

Thirdly, as another limitation, there is no direct connection between the modeling results generated by ML or NLP and the real functioning of a PED. Also, the characterization of a PED is still often based on only a single element. However, factors such as policy, social acceptance, and economic viability also play a crucial role in the success of PEDs. A holistic approach to evaluating PEDs, where ML and NLP can be assessed and integrated alongside

other aspects of PED planning and implementation, is necessary if the desired outcomes are to be achieved. This holistic approach will require a multidisciplinary knowledge that considers the multiple aspects of PEDs, as well as an awareness of the technological and computational requirements. The integration of more transparent model architectures and the development of user-friendly interfaces that can present model insights in an understandable manner need to be considered. The process will also involve engaging with stakeholders and policymakers to ensure that the outcomes of ML and NLP models align with the broader goals and objectives of PEDs.

Lastly, there is a real shortage of widely applicable conclusions to be drawn from the ML and NLP analysis of PED elements to date. Right now, it is very difficult to compare PED studies with each other because of the widely different factors affecting each one, such as input data, output data, time and space resolution, and location. This makes it hard to draw firm conclusions about the generalizability of AI techniques. It also limits the generalizability of PED characterizations that can be drawn from these models. In the future, these well-tuned models could benefit from a greater focus on feature selection if more general frameworks for PEDs are to be developed and utilized. It also involves synthesizing the insights gained from these models with other relevant information to inform decision-making processes. For example, generalizing ML and NLP findings by using geographic information systems (GIS) and demographic data ensures that the energy solutions are tailored to the specific characteristics of the district. Existing urban infrastructure data can also be integrated to identify areas that need upgrades. Successful implementation of these initiatives necessitates collaborative efforts and a shared understanding between districts and local governments to ensure alignment with PED plans.

6. Conclusions

This paper reviews machine learning and natural language processing methods for characterizing PEDs. Nineteen elements representing both energy and non-energy aspects of PEDs were identified from the literature. Approximately seventy research papers were reviewed to gain a deeper understanding of how ML and NLP methods could be used to model these elements. The most important models for ML are ANN, SVM, and tree-based approaches that can be used for prediction and classification at different phases. For NLP, tasks like topic modeling, word embedding, and semantic modeling using large language models like BERT have been found to be effective for stakeholder matching, sentiment analysis, and metadata analysis. However, a number of elements of a PED, such as energy flexibility, the energy carrier, interaction with the grid, governance, spatial resolution, and justice, have not shown any potential to be modeled by ML or NLP. This gap pinpoints the need for research to explore and develop methodologies capable of addressing these aspects.

By adding the results that ML and NLP models are able to achieve to a more holistic analysis of the comparative performance of PEDs using other possible measures, this approach ensures that the vision of creating sustainable and resilient urban environments is much closer to being realized. This includes comparing their performance against other measures, defining and quantifying the PED elements, utilizing techniques in urban-scale energy modeling, and assessing their effectiveness and transparency in a broader and user-friendly context. Future studies should focus on continuously validating these methods, not only to establish their reliability but also to enhance their capability in the intricate AI tasks associated with PEDs.

References

1. 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. [CrossRef]
2. Koutra, S.; Terés-Zubiaga, J.; Bouillard, P.; Becue, V. 'Decarbonizing Europe' A Critical Review on Positive Energy Districts Approaches. *Sustain. Cities Soc.* **2023**, *89*, 104356. [CrossRef]
3. *SET-Plan ACTION N°3.2 Implementation Plan*; European Commission: Brussels, Belgium, 2018.
4. Bossi, S.; Gollner, C.; Theierling, S. Towards 100 Positive Energy Districts in Europe: Preliminary Data Analysis of 61 European Cases. *Energies* **2020**, *13*, 6083. [CrossRef]
5. Albert-Seifried, V.; Murauskaite, L.; Massa, G.; Aelenei, L.; Baer, D.; Krangsås, S.G.; Alpagut, B.; Mutule, A.; Pokorny, N.; Vandevyvere, H. Definitions of Positive Energy Districts: A Review of the Status Quo and Challenges. In *Sustainability in Energy and Buildings 2021*; Littlewood, J.R., Howlett, R.J., Jain, L.C., Eds.; Smart Innovation, Systems and Technologies; Springer: Singapore, 2022; Volume 263, pp. 493–506, ISBN 9789811662683.
6. Cheng, C.; Albert-Seifried, V.; Aelenei, L.; Vandevyvere, H.; Seco, O.; Nuria Sánchez, M.; Hukkalainen, M. A Systematic Approach Towards Mapping Stakeholders in Different Phases of PED Development—Extending the PED Toolbox. In *Sustainability in Energy and Buildings 2021*; Littlewood, J.R., Howlett, R.J., Jain, L.C., Eds.; Smart Innovation, Systems and Technologies; Springer: Singapore, 2022; Volume 263, pp. 447–463, ISBN 9789811662683.
7. Belda, A.; Giancola, E.; Williams, K.; Dabirian, S.; Jradi, M.; Volpe, R.; Abolhassani, S.S.; Fichera, A.; Eicker, U. Reviewing Challenges and Limitations of Energy Modelling Software in the Assessment of PEDs Using Case Studies. In *Sustainability in Energy and Buildings 2021*; Littlewood, J.R., Howlett, R.J., Jain, L.C., Eds.; Smart Innovation, Systems and Technologies; Springer: Singapore, 2022; Volume 263, pp. 465–477, ISBN 9789811662683.
8. Steemers, K.; Gohari Krangsas, S.; Ashrafan, T.; Giancola, E.; Konstantinou, T.; Liu, M.; Maas, N.; Murauskaite, L.; Prebreza, B.; Soutullo, S. Challenges for a Positive Energy District Framework. *Sustain. Energy Build. Res. Adv.* **2021**, *8*, 10–19. [CrossRef]
9. Fatima, Z.; Padilla, M.; Kuzmic, M.; Huovila, A.; Schaj, G.; Effenberger, N. Positive Energy Districts: The 10 Replicated Solutions in Maia, Reykjavik, Kifissia, Kladno and Lviv. *Smart Cities* **2022**, *6*, 1–18. [CrossRef]
10. Leone, F.; Reda, F.; Hasan, A.; Rehman, H.U.; Nigrelli, F.C.; Nocera, F.; Costanzo, V. Lessons Learned from Positive Energy District (PED) Projects: Cataloguing and Analysing Technology Solutions in Different Geographical Areas in Europe. *Energies* **2022**, *16*, 356. [CrossRef]
11. Shnapp, S.; Paci, D.; Bertoldi, P. *Enabling Positive Energy Districts across Europe: Energy Efficiency Couples Renewable Energy*; Publications Office of the European Union: Luxembourg, 2020; ISBN 978-92-76-21043-6.
12. Derkenbaeva, E.; Yoo, H.K.; Hofstede, G.J.; Galanakis, K.; Ackrill, R. Positive Energy Districts in Europe: One Size Does Not Fit All. In Proceedings of the 2022 IEEE International Smart Cities Conference (ISC2), Pafos, Cyprus, 26 September 2022; pp. 1–6.
13. Bruck, A.; Díaz Ruano, S.; Auer, H. One Piece of the Puzzle towards 100 Positive Energy Districts (PEDs) across Europe by 2025: An Open-Source Approach to Unveil Favourable Locations of PV-Based PEDs from a Techno-Economic Perspective. *Energy* **2022**, *254*, 124152. [CrossRef]
14. Rueda Castellanos, S.; Oregi, X. Positive Energy District (PED) Selected Projects Assessment, Study towards the Development of Further PEDs. *Environ. Clim. Technol.* **2021**, *25*, 281–294. [CrossRef]
15. 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. [CrossRef]
16. Lindholm, O.; Rehman, H.U.; Reda, F. Positioning Positive Energy Districts in European Cities. *Buildings* **2021**, *11*, 19. [CrossRef]
17. 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. [CrossRef]
18. Alpagut, B.; Akyürek, Ö.; Mitre, E.M. Positive Energy Districts Methodology and Its Replication Potential. In Proceedings of the Sustainable Places 2019, Cagliari, Italy, 5–7 June 2019; p. 8.
19. Bambara, J.; Athienitis, A.K.; Eicker, U. Residential Densification for Positive Energy Districts. *Front. Sustain. Cities* **2021**, *3*, 630973. [CrossRef]

20. Zhou, Y.; Cao, S.; Hensen, J.L.M. An Energy Paradigm Transition Framework from Negative towards Positive District Energy Sharing Networks—Battery Cycling Aging, Advanced Battery Management Strategies, Flexible Vehicles-to-Buildings Interactions, Uncertainty and Sensitivity Analysis. *Appl. Energy* **2021**, *288*, 116606. [CrossRef]

21. Civiero, P.; Pascual, J.; Arcas Abella, J.; Bilbao Figuero, A.; Salom, J. PEDRERA. Positive Energy District Renovation Model for Large Scale Actions. *Energies* **2021**, *14*, 2833. [CrossRef]

22. Soutullo, S.; Aelenei, L.; Nielsen, P.S.; Ferrer, J.A.; Gonçalves, H. Testing Platforms as Drivers for Positive-Energy Living Laboratories. *Energies* **2020**, *13*, 5621. [CrossRef]

23. Fatima, Z.; Pollmer, U.; Santala, S.-S.; Kontu, K.; Ticklen, M. Citizens and Positive Energy Districts: Are Espoo and Leipzig Ready for PEDs? *Buildings* **2021**, *11*, 102. [CrossRef]

24. Baer, D.; Loewen, B.; Cheng, C.; Thomsen, J.; Wyckmans, A.; Temeljotov-Salaj, A.; Ahlers, D. Approaches to Social Innovation in Positive Energy Districts (PEDs)—A Comparison of Norwegian Projects. *Sustainability* **2021**, *13*, 7362. [CrossRef]

25. Zhou, Y. Advances of Machine Learning in Multi-Energy District Communities—Mechanisms, Applications and Perspectives. *Energy AI* **2022**, *10*, 100187. [CrossRef]

26. Zhang, X.; Shen, J.; Saini, P.K.; Lovati, M.; Han, M.; Huang, P.; Huang, Z. Digital Twin for Accelerating Sustainability in Positive Energy District: A Review of Simulation Tools and Applications. *Front. Sustain. Cities* **2021**, *3*, 663269. [CrossRef]

27. Sareen, S.; Albert-Seifried, V.; Aelenei, L.; Reda, F.; Etminan, G.; Andreucci, M.-B.; Kuzmic, M.; Maas, N.; Seco, O.; Civiero, P.; et al. Ten Questions Concerning Positive Energy Districts. *Build. Environ.* **2022**, *216*, 109017. [CrossRef]

28. Wei, Y.; Zhang, X.; Shi, Y.; Xia, L.; Pan, S.; Wu, J.; Han, M.; Zhao, X. A Review of Data-Driven Approaches for Prediction and Classification of Building Energy Consumption. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1027–1047. [CrossRef]

29. Deb, C.; Schlueter, A. Review of Data-Driven Energy Modelling Techniques for Building Retrofit. *Renew. Sustain. Energy Rev.* **2021**, *144*, 110990. [CrossRef]

30. Jordan, M.I.; Mitchell, T.M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]

31. Hinterberger, R.; Gollne, C.; Noll, M.; Meyer, S.; Schwarz, H.-G. *White Paper on PED Reference Framework for Positive Energy Districts and Neighbourhoods*; Austrian Research Promotion Agency: Wien, Austria, 2020.

32. Casamassima, L.; Bottecchia, L.; Bruck, A.; Kranzl, L.; Haas, R. Economic, Social, and Environmental Aspects of Positive Energy Districts—A Review. *WIREs Energy Environ.* **2022**, *11*, e452. [CrossRef]

33. Cellura, M.; Fichera, A.; Guarino, F.; Volpe, R. Sustainable Development Goals and Performance Measurement of Positive Energy District: A Methodological Approach. In *Sustainability in Energy and Buildings 2021*; Littlewood, J.R., Howlett, R.J., Jain, L.C., Eds.; Smart Innovation, Systems and Technologies; Springer: Singapore, 2022; Volume 263, pp. 519–527, ISBN 9789811662683.

34. Guarino, F.; Bisello, A.; Frieden, D.; Bastos, J.; Brunetti, A.; Cellura, M.; Ferraro, M.; Fichera, A.; Giancola, E.; Haase, M.; et al. State of the Art on Sustainability Assessment of Positive Energy Districts: Methodologies, Indicators and Future Perspectives. In *Sustainability in Energy and Buildings 2021*; Littlewood, J.R., Howlett, R.J., Jain, L.C., Eds.; Smart Innovation, Systems and Technologies; Springer: Singapore, 2022; Volume 263, pp. 479–492, ISBN 9789811662683.

35. 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, 2022; Volume 263, pp. 507–517, ISBN 9789811662683.

36. Bottecchia, L.; Gabaldón, A.; Castillo-Calzadilla, T.; Soutullo, S.; Ranjbar, S.; Eicker, U. Fundamentals of Energy Modelling for Positive Energy Districts. In *Sustainability in Energy and Buildings 2021*; Littlewood, J.R., Howlett, R.J., Jain, L.C., Eds.; Smart Innovation, Systems and Technologies; Springer: Singapore, 2022; Volume 263, pp. 435–445, ISBN 9789811662683.

37. Derkenbaeva, E.; Halleck Vega, S.; Hofstede, G.J.; van Leeuwen, E. Positive Energy Districts: Mainstreaming Energy Transition in Urban Areas. *Renew. Sustain. Energy Rev.* **2022**, *153*, 111782. [CrossRef]

38. Ngarambe, J.; Yun, G.Y.; Santamouris, M. The Use of Artificial Intelligence (AI) Methods in the Prediction of Thermal Comfort in Buildings: Energy Implications of AI-Based Thermal Comfort Controls. *Energy Build.* **2020**, *211*, 109807. [CrossRef]

39. Wang, Z.; Srinivasan, R.S. A Review of Artificial Intelligence Based Building Energy Use Prediction: Contrasting the Capabilities of Single and Ensemble Prediction Models. *Renew. Sustain. Energy Rev.* **2017**, *75*, 796–808. [CrossRef]

40. Halhoul Merabet, G.; Essaaidi, M.; Ben Haddou, M.; Qolomany, B.; Qadir, J.; Anan, M.; Al-Fuqaha, A.; Abid, M.R.; Benhaddou, D. Intelligent Building Control Systems for Thermal Comfort and Energy-Efficiency: A Systematic Review of Artificial Intelligence-Assisted Techniques. *Renew. Sustain. Energy Rev.* **2021**, *144*, 110969. [CrossRef]

41. Zhao, Y.; Li, T.; Zhang, X.; Zhang, C. Artificial Intelligence-Based Fault Detection and Diagnosis Methods for Building Energy Systems: Advantages, Challenges and the Future. *Renew. Sustain. Energy Rev.* **2019**, *109*, 85–101. [CrossRef]

42. Sacks, R.; Wang, Z.; Ouyang, B.; Utkucu, D.; Chen, S. Toward Artificially Intelligent Cloud-Based Building Information Modelling for Collaborative Multidisciplinary Design. *Adv. Eng. Inform.* **2022**, *53*, 101711. [CrossRef]

43. Ahmed, W.; Ansari, H.; Khan, B.; Ullah, Z.; Ali, S.M.; Mehmood, C.A.A.; Qureshi, M.B.; Hussain, I.; Jawad, M.; Khan, M.U.S.; et al. Machine Learning Based Energy Management Model for Smart Grid and Renewable Energy Districts. *IEEE Access* **2020**, *8*, 185059–185078. [CrossRef]

44. Ntakolia, C.; Anagnostis, A.; Moustakidis, S.; Karcanias, N. Machine Learning Applied on the District Heating and Cooling Sector: A Review. *Energy Syst.* **2022**, *13*, 1–30. [CrossRef]

45. Husiev, O.; Ukar Arrien, O.; Enciso-Santocildes, M. What Does Horizon 2020 Contribute to? Analysing and Visualising the Community Practices of Europe's Largest Research and Innovation Programme. *Energy Res. Soc. Sci.* **2023**, *95*, 102879. [CrossRef]
46. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805. [CrossRef]
47. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. *arXiv* **2018**, arXiv:1802.05365. [CrossRef]
48. Hong, T.; Chen, Y.; Luo, X.; Luo, N.; Lee, S.H. Ten Questions on Urban Building Energy Modeling. *Build. Environ.* **2020**, *168*, 106508. [CrossRef]
49. Österbring, M.; Mata, É.; Thuvander, L.; Mangold, M.; Johnsson, F.; Wallbaum, H. A Differentiated Description of Building-Stocks for a Georeferenced Urban Bottom-up Building-Stock Model. *Energy Build.* **2016**, *120*, 78–84. [CrossRef]
50. Goy, S.; Maréchal, F.; Finn, D. Data for Urban Scale Building Energy Modelling: Assessing Impacts and Overcoming Availability Challenges. *Energies* **2020**, *13*, 4244. [CrossRef]
51. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer Science & Business Media, LLC: New York, NY, USA, 2006.
52. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; Adaptive Computation and Machine Learning Series; The MIT Press: Cambridge, MA, USA, 2018; ISBN 978-0-262-03924-6.
53. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*, 2nd ed.; Adaptive Computation and Machine Learning; The MIT Press: Cambridge, MA, USA, 2018; ISBN 978-0-262-03940-6.
54. Foucquier, A.; Robert, S.; Suard, F.; Stéphan, L.; Jay, A. State of the Art in Building Modelling and Energy Performances Prediction: A Review. *Renew. Sustain. Energy Rev.* **2013**, *23*, 272–288. [CrossRef]
55. Amasyali, K.; El-Gohary, N.M. A Review of Data-Driven Building Energy Consumption Prediction Studies. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1192–1205. [CrossRef]
56. Alasadi, S.A.; Bhaya, W.S. Review of Data Preprocessing Techniques in Data Mining. *J. Eng. Appl. Sci.* **2017**, *12*, 4102–4107.
57. Ayoub, M. A Review on Machine Learning Algorithms to Predict Daylighting inside Buildings. *Sol. Energy* **2020**, *202*, 249–275. [CrossRef]
58. Wu, J.; Chen, X.-Y.; Zhang, H.; Xiong, L.-D.; Lei, H.; Deng, S.-H. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *J. Electron. Sci. Technol.* **2019**, *17*, 26–40.
59. Seyedzadeh, S.; Rahimian, F.P.; Glesk, I.; Roper, M. Machine Learning for Estimation of Building Energy Consumption and Performance: A Review. *Vis. Eng.* **2018**, *6*, 5. [CrossRef]
60. Conforti, C.; Hirmer, S.; Morgan, D.; Basaldella, M.; Or, Y.B. Natural Language Processing for Achieving Sustainable Development: The Case of Neural Labelling to Enhance Community Profiling. *arXiv* **2020**, arXiv:2004.12935. [CrossRef]
61. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K.A.; Ceder, G.; Jain, A. Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* **2019**, *571*, 95–98. [CrossRef] [PubMed]
62. Zekić-Sušac, M.; Mitrović, S.; Has, A. Machine Learning Based System for Managing Energy Efficiency of Public Sector as an Approach towards Smart Cities. *Int. J. Inf. Manag.* **2021**, *58*, 102074. [CrossRef]
63. Narciso, D.A.C.; Martins, F.G. Application of Machine Learning Tools for Energy Efficiency in Industry: A Review. *Energy Rep.* **2020**, *6*, 1181–1199. [CrossRef]
64. Zhang, W.; Wu, Y.; Calautit, J.K. A Review on Occupancy Prediction through Machine Learning for Enhancing Energy Efficiency, Air Quality and Thermal Comfort in the Built Environment. *Renew. Sustain. Energy Rev.* **2022**, *167*, 112704. [CrossRef]
65. Wang, Z.; Liu, J.; Zhang, Y.; Yuan, H.; Zhang, R.; Srinivasan, R.S. Practical Issues in Implementing Machine-Learning Models for Building Energy Efficiency: Moving beyond Obstacles. *Renew. Sustain. Energy Rev.* **2021**, *143*, 110929. [CrossRef]
66. Said, O.; Tolba, A. Accurate Performance Prediction of IoT Communication Systems for Smart Cities: An Efficient Deep Learning Based Solution. *Sustain. Cities Soc.* **2021**, *69*, 102830. [CrossRef]
67. Ullah, Z.; Al-Turjman, F.; Mostarda, L.; Gagliardi, R. Applications of Artificial Intelligence and Machine Learning in Smart Cities. *Comput. Commun.* **2020**, *154*, 313–323. [CrossRef]
68. Bhattacharya, S.; Somayaji, S.R.K.; Gadekallu, T.R.; Alazab, M.; Maddikunta, P.K.R. A Review on Deep Learning for Future Smart Cities. *Internet Technol. Lett.* **2022**, *5*, e187. [CrossRef]
69. Serrano, W. Deep Reinforcement Learning Algorithms in Intelligent Infrastructure. *Infrastructures* **2019**, *4*, 52. [CrossRef]
70. Benavente-Peces, C.; Ibadah, N. Buildings Energy Efficiency Analysis and Classification Using Various Machine Learning Technique Classifiers. *Energies* **2020**, *13*, 3497. [CrossRef]
71. Olu-Ajayi, R.; Alaka, H.; Sulaimon, I.; Sunmola, F.; Ajayi, S. Building Energy Consumption Prediction for Residential Buildings Using Deep Learning and Other Machine Learning Techniques. *J. Build. Eng.* **2022**, *45*, 103406. [CrossRef]
72. Shapi, M.K.M.; Ramli, N.A.; Awalin, L.J. Energy Consumption Prediction by Using Machine Learning for Smart Building: Case Study in Malaysia. *Dev. Built Environ.* **2021**, *5*, 100037. [CrossRef]
73. Khalil, M.; McGough, A.S.; Pourmirza, Z.; Pazhoohesh, M.; Walker, S. Machine Learning, Deep Learning and Statistical Analysis for Forecasting Building Energy Consumption—A Systematic Review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105287. [CrossRef]
74. Liu, Z.; Wu, D.; Liu, Y.; Han, Z.; Lun, L.; Gao, J.; Jin, G.; Cao, G. Accuracy Analyses and Model Comparison of Machine Learning Adopted in Building Energy Consumption Prediction. *Energy Explor. Exploit.* **2019**, *37*, 1426–1451. [CrossRef]
75. Wang, Z.; Hong, T.; Piette, M.A. Building Thermal Load Prediction through Shallow Machine Learning and Deep Learning. *Appl. Energy* **2020**, *263*, 114683. [CrossRef]

76. Forouzandeh, N.; Zomorodian, Z.S.; Shaghaghian, Z.; Tahsildoost, M. Room Energy Demand and Thermal Comfort Predictions in Early Stages of Design Based on the Machine Learning Methods. *Intell. Build. Int.* **2022**, *15*, 3–20. [CrossRef]

77. Fathi, S.; Srinivasan, R.; Fenner, A.; Fathi, S. Machine Learning Applications in Urban Building Energy Performance Forecasting: A Systematic Review. *Renew. Sustain. Energy Rev.* **2020**, *133*, 110287. [CrossRef]

78. Ahmadiahangar, R.; Haring, T.; Rosin, A.; Korotko, T.; Martins, J. Residential Load Forecasting for Flexibility Prediction Using Machine Learning-Based Regression Model. In Proceedings of the 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), Genova, Italy, 11–14 June 2019; pp. 1–4.

79. Kim, S.-G.; Jung, J.-Y.; Sim, M. A Two-Step Approach to Solar Power Generation Prediction Based on Weather Data Using Machine Learning. *Sustainability* **2019**, *11*, 1501. [CrossRef]

80. Avila, D.; Marichal, G.N.; Padrón, I.; Quiza, R.; Hernández, Á. Forecasting of Wave Energy in Canary Islands Based on Artificial Intelligence. *Appl. Ocean Res.* **2020**, *101*, 102189. [CrossRef]

81. Zhang, Z.; Ye, L.; Qin, H.; Liu, Y.; Wang, C.; Yu, X.; Yin, X.; Li, J. Wind Speed Prediction Method Using Shared Weight Long Short-Term Memory Network and Gaussian Process Regression. *Appl. Energy* **2019**, *247*, 270–284. [CrossRef]

82. Zhang, X.; Peng, Y.; Xu, W.; Wang, B. An Optimal Operation Model for Hydropower Stations Considering Inflow Forecasts with Different Lead-Times. *Water Resour. Manag.* **2019**, *33*, 173–188. [CrossRef]

83. khan, M.; Raza Naqvi, S.; Ullah, Z.; Ali Ammar Taqvi, S.; Nouman Aslam Khan, M.; Farooq, W.; Taqi Mehran, M.; Juchelková, D.; Štěpanec, L. Applications of Machine Learning in Thermochemical Conversion of Biomass-A Review. *Fuel* **2023**, *332*, 126055. [CrossRef]

84. Okoroafor, E.R.; Smith, C.M.; Ochie, K.I.; Nwosu, C.J.; Gudmundsdottir, H.; Aljubran, M.J. Machine Learning in Subsurface Geothermal Energy: Two Decades in Review. *Geothermics* **2022**, *102*, 102401. [CrossRef]

85. Khan, P.W.; Byun, Y.-C.; Lee, S.-J.; Kang, D.-H.; Kang, J.-Y.; Park, H.-S. Machine Learning-Based Approach to Predict Energy Consumption of Renewable and Nonrenewable Power Sources. *Energies* **2020**, *13*, 4870. [CrossRef]

86. Abd El-Aziz, R.M. Renewable Power Source Energy Consumption by Hybrid Machine Learning Model. *Alex. Eng. J.* **2022**, *61*, 9447–9455. [CrossRef]

87. Aslam, M.; Lee, J.-M.; Kim, H.-S.; Lee, S.-J.; Hong, S. Deep Learning Models for Long-Term Solar Radiation Forecasting Considering Microgrid Installation: A Comparative Study. *Energies* **2019**, *13*, 147. [CrossRef]

88. Malakouti, S.M. Use Machine Learning Algorithms to Predict Turbine Power Generation to Replace Renewable Energy with Fossil Fuels. *Energy Explor. Exploit.* **2023**, *41*, 836–857. [CrossRef]

89. Rohani, A.; Taki, M.; Abdollahpour, M. A Novel Soft Computing Model (Gaussian Process Regression with K-Fold Cross Validation) for Daily and Monthly Solar Radiation Forecasting (Part: I). *Renew. Energy* **2018**, *115*, 411–422. [CrossRef]

90. Kaab, A.; Sharifi, M.; Mobli, H.; Nabavi-Pelesaraei, A.; Chau, K. Combined Life Cycle Assessment and Artificial Intelligence for Prediction of Output Energy and Environmental Impacts of Sugarcane Production. *Sci. Total Environ.* **2019**, *664*, 1005–1019. [CrossRef] [PubMed]

91. Nilashi, M.; Rupani, P.F.; Rupani, M.M.; Kamyab, H.; Shao, W.; Ahmadi, H.; Rashid, T.A.; Aljojo, N. Measuring Sustainability through Ecological Sustainability and Human Sustainability: A Machine Learning Approach. *J. Clean. Prod.* **2019**, *240*, 118162. [CrossRef]

92. Valladares, W.; Galindo, M.; Gutiérrez, J.; Wu, W.-C.; Liao, K.-K.; Liao, J.-C.; Lu, K.-C.; Wang, C.-C. Energy Optimization Associated with Thermal Comfort and Indoor Air Control via a Deep Reinforcement Learning Algorithm. *Build. Environ.* **2019**, *155*, 105–117. [CrossRef]

93. Kim, J.; Zhou, Y.; Schiavon, S.; Raftery, P.; Brager, G. Personal Comfort Models: Predicting Individuals' Thermal Preference Using Occupant Heating and Cooling Behavior and Machine Learning. *Build. Environ.* **2018**, *129*, 96–106. [CrossRef]

94. Pinto, G.; Deltetto, D.; Capozzoli, A. Data-Driven District Energy Management with Surrogate Models and Deep Reinforcement Learning. *Appl. Energy* **2021**, *304*, 117642. [CrossRef]

95. Han, M.; May, R.; Zhang, X.; Wang, X.; Pan, S.; Yan, D.; Jin, Y.; Xu, L. A Review of Reinforcement Learning Methodologies for Controlling Occupant Comfort in Buildings. *Sustain. Cities Soc.* **2019**, *51*, 101748. [CrossRef]

96. Das, H.S.; Roy, P. A Deep Dive Into Deep Learning Techniques for Solving Spoken Language Identification Problems. In *Intelligent Speech Signal Processing*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 81–100, ISBN 978-0-12-818130-0.

97. Olu-Ajayi, R.; Alaka, H.; Sulaimon, I.; Sunmola, F.; Ajayi, S. Machine Learning for Energy Performance Prediction at the Design Stage of Buildings. *Energy Sustain. Dev.* **2022**, *66*, 12–25. [CrossRef]

98. Wang, H.; Lei, Z.; Zhang, X.; Zhou, B.; Peng, J. A Review of Deep Learning for Renewable Energy Forecasting. *Energy Convers. Manag.* **2019**, *198*, 111799. [CrossRef]

99. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]

100. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

101. Ma, J.; Cheng, J.C.P. Identifying the Influential Features on the Regional Energy Use Intensity of Residential Buildings Based on Random Forests. *Appl. Energy* **2016**, *183*, 193–201. [CrossRef]

102. Lai, J.-P.; Chang, Y.-M.; Chen, C.-H.; Pai, P.-F. A Survey of Machine Learning Models in Renewable Energy Predictions. *Appl. Sci.* **2020**, *10*, 5975. [CrossRef]

103. De Las Heras, A.; Luque-Sendra, A.; Zamora-Polo, F. Machine Learning Technologies for Sustainability in Smart Cities in the Post-COVID Era. *Sustainability* **2020**, *12*, 9320. [CrossRef]

104. Rangel-Martinez, D.; Nigam, K.D.P.; Ricardez-Sandoval, L.A. Machine Learning on Sustainable Energy: A Review and Outlook on Renewable Energy Systems, Catalysis, Smart Grid and Energy Storage. *Chem. Eng. Res. Des.* **2021**, *174*, 414–441. [CrossRef]

105. 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. [CrossRef]

106. Khanuja, A.; Webb, A.L. What We Talk about When We Talk about EEMs: Using Text Mining and Topic Modeling to Understand Building Energy Efficiency Measures (1836-RP). *Sci. Technol. Built Environ.* **2023**, *29*, 4–18. [CrossRef]

107. Chen, W.; Zhou, Y.; Wu, Q.; Chen, G.; Huang, X.; Yu, B. Urban Building Type Mapping Using Geospatial Data: A Case Study of Beijing, China. *Remote Sens.* **2020**, *12*, 2805. [CrossRef]

108. Xue, X.; Zhang, J. Part-of-Speech Tagging of Building Codes Empowered by Deep Learning and Transformational Rules. *Adv. Eng. Inform.* **2021**, *47*, 101235. [CrossRef]

109. Ding, Z.; Li, Z.; Fan, C. Building Energy Savings: Analysis of Research Trends Based on Text Mining. *Autom. Constr.* **2018**, *96*, 398–410. [CrossRef]

110. Bickel, M.W. Reflecting Trends in the Academic Landscape of Sustainable Energy Using Probabilistic Topic Modeling. *Energy Sustain. Soc.* **2019**, *9*, 49. [CrossRef]

111. 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. [CrossRef]

112. Saheb, T.; Dehghani, M.; Saheb, T. Artificial Intelligence for Sustainable Energy: A Contextual Topic Modeling and Content Analysis. *Sustain. Comput. Inform. Syst.* **2022**, *35*, 100699. [CrossRef]

113. Wu, Z.; He, Q.; Chen, Q.; Xue, H.; Li, S. A Topical Network Based Analysis and Visualization of Global Research Trends on Green Building from 1990 to 2020. *J. Clean. Prod.* **2021**, *320*, 128818. [CrossRef]

114. 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 Build.* **2021**, *242*, 110885. [CrossRef]

115. Wang, B.; Wang, Z. Heterogeneity Evaluation of China's Provincial Energy Technology Based on Large-Scale Technical Text Data Mining. *J. Clean. Prod.* **2018**, *202*, 946–958. [CrossRef]

116. Lai, Y.; Papadopoulos, S.; Fuerst, F.; Pivo, G.; Sagi, J.; Kontokosta, C.E. Building Retrofit Hurdle Rates and Risk Aversion in Energy Efficiency Investments. *Appl. Energy* **2022**, *306*, 118048. [CrossRef]

117. Copiello, S. Economic Parameters in the Evaluation Studies Focusing on Building Energy Efficiency: A Review of the Underlying Rationale, Data Sources, and Assumptions. *Energy Procedia* **2019**, *157*, 180–192. [CrossRef]

118. Zeng, R.; Chini, A. A Review of Research on Embodied Energy of Buildings Using Bibliometric Analysis. *Energy Build.* **2017**, *155*, 172–184. [CrossRef]

119. Jung, N.; Lee, G. Automated Classification of Building Information Modeling (BIM) Case Studies by BIM Use Based on Natural Language Processing (NLP) and Unsupervised Learning. *Adv. Eng. Inform.* **2019**, *41*, 100917. [CrossRef]

120. Waterworth, D.; Sethuvenkatraman, S.; Sheng, Q.Z. Advancing Smart Building Readiness: Automated Metadata Extraction Using Neural Language Processing Methods. *Adv. Appl. Energy* **2021**, *3*, 100041. [CrossRef]

121. Xue, X.; Zhang, J. Building Codes Part-of-Speech Tagging Performance Improvement by Error-Driven Transformational Rules. *J. Comput. Civ. Eng.* **2020**, *34*, 04020035. [CrossRef]

122. Jiao, Y.; Li, J.; Wu, J.; Hong, D.; Gupta, R.; Shang, J. SeNsER: Learning Cross-Building Sensor Metadata Tagger. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 950–960.

123. Zhan, S.; Chong, A.; Lasternas, B. Automated Recognition and Mapping of Building Management System (BMS) Data Points for Building Energy Modeling (BEM). *Build. Simul.* **2021**, *14*, 43–52. [CrossRef]

124. Stinner, F.; Neißer-Deiters, P.; Baranski, M.; Müller, D. Aikido: Structuring Data Point Identifiers of Technical Building Equipment by Machine Learning. *J. Phys. Conf. Ser.* **2019**, *1343*, 012039. [CrossRef]

125. Nalmpantis, C.; Krystalakos, O.; Vrakas, D. Energy Profile Representation in Vector Space. In Proceedings of the 10th Hellenic Conference on Artificial Intelligence, Patras, Greece, 9–12 July 2018; pp. 1–5.

126. 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 Build.* **2021**, *249*, 111211. [CrossRef]

127. Nie, Z.; Yang, Y.; Xu, Q. An Ensemble-Policy Non-Intrusive Load Monitoring Technique Based Entirely on Deep Feature-Guided Attention Mechanism. *Energy Build.* **2022**, *273*, 112356. [CrossRef]

128. Abdelaziz, A.; Santos, V.; Dias, M.S. Machine Learning Techniques in the Energy Consumption of Buildings: A Systematic Literature Review Using Text Mining and Bibliometric Analysis. *Energies* **2021**, *14*, 7810. [CrossRef]

129. Kumar, M.; Ng, J. Using Text Mining and Topic Modelling to Understand Success and Growth Factors in Global Renewable Energy Projects. *Renew. Energy Focus* **2022**, *42*, 211–220. [CrossRef]

130. 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. [CrossRef]

131. 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. [CrossRef]

132. 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. [CrossRef] [PubMed]

133. Jain, A.; Jain, V. Sentiment Classification of Twitter Data Belonging to Renewable Energy Using Machine Learning. *J. Inf. Optim. Sci.* **2019**, *40*, 521–533. [CrossRef]

134. 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. [CrossRef] [PubMed]

135. Tie, M.; Zhu, M. Interpreting Low-Carbon Transition at the Subnational Level: Evidence from China Using a Natural Language Processing Approach. *Resour. Conserv. Recycl.* **2022**, *187*, 106636. [CrossRef]

136. Nicolas, C.; Kim, J.; Chi, S. Natural Language Processing-Based Characterization of Top-down Communication in Smart Cities for Enhancing Citizen Alignment. *Sustain Cities Soc.* **2021**, *66*, 102674. [CrossRef]

137. Cai, R.; Qin, B.; Chen, Y.; Zhang, L.; Yang, R.; Chen, S.; Wang, W. Sentiment Analysis About Investors and Consumers in Energy Market Based on BERT-BiLSTM. *IEEE Access* **2020**, *8*, 171408–171415. [CrossRef]

138. Yao, Y.; Li, X.; Liu, X.; Liu, P.; Liang, Z.; Zhang, J.; Mai, K. Sensing Spatial Distribution of Urban Land Use by Integrating Points-of-Interest and Google Word2Vec Model. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 825–848. [CrossRef]

139. Zhai, W.; Bai, X.; Shi, Y.; Han, Y.; Peng, Z.-R.; Gu, C. Beyond Word2vec: An Approach for Urban Functional Region Extraction and Identification by Combining Place2vec and POIs. *Comput. Environ. Urban Syst.* **2019**, *74*, 1–12. [CrossRef]

140. Zhong, Y.; Su, Y.; Wu, S.; Zheng, Z.; Zhao, J.; Ma, A.; Zhu, Q.; Ye, R.; Li, X.; Pellikka, P.; et al. Open-Source Data-Driven Urban Land-Use Mapping Integrating Point-Line-Polygon Semantic Objects: A Case Study of Chinese Cities. *Remote Sens. Environ.* **2020**, *247*, 111838. [CrossRef]

141. Yuan, J.; Zheng, Y.; Xie, X. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 186–194.

142. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.

143. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. *arXiv* **2013**, arXiv:1310.4546. [CrossRef]

144. Rong, X. Word2vec Parameter Learning Explained. *arXiv* **2016**, arXiv:1411.2738.

145. Blei, D.M.; NG, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.

146. Blei, D.M. Probabilistic Topic Models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]

147. Asuncion, A.; Welling, M.; Smyth, P.; Teh, Y.W. On Smoothing and Inference for Topic Models. *arXiv* **2012**, arXiv:1205.2662. [CrossRef]

148. *ASHRAE Standard 100-2018*; Energy Efficiency in Existing Buildings. ASHRAE: Atlanta, GA, USA, 2018.

149. Zhivov, A.; Cyrus, N. Energy Efficient Technologies and Measures for Building Renovation: Sourcebook. IEA ECBS Annex 46. 2014. Available online: https://www.iea-ebc.org/Data/publications/EBC_Annex_46_Technologies_and_Measures_Sourcebook.pdf (accessed on 25 January 2024).

150. Doty, S. *Commercial Energy Auditing Reference Handbook*, 2nd ed.; Fairmont Press: Boca Raton, FL, USA, 2011.

151. National Renewable Energy Laboratory. National Residential Energy Efficiency Measures Database, Version 3.1.0. 2018. Available online: https://remdb.nrel.gov/ (accessed on 25 January 2024).

152. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]

153. Neumann, H.-M.; Hainoun, A.; Stollnberger, R.; Etminan, G.; Schaffler, V. Analysis and Evaluation of the Feasibility of Positive Energy Districts in Selected Urban Typologies in Vienna Using a Bottom-Up District Energy Modelling Approach. *Energies* **2021**, *14*, 4449. [CrossRef]