Scientometrics
https://doi.org/10.1007/s11192-025-05394-z

Field identification and opportunity discovery of photovoltaics technology: deep transfer learning method

Ruilian Han[1,2] · Lu An[1,2,3] · Wei Zhou[2] · Gang Li[1,2]

Received: 18 June 2024 / Accepted: 22 July 2025

Abstract
The advancement of technology relies on scientific and accurate identification of potential opportunities within the field. This study presents a novel method for discovering technology opportunities by combining multi-source data sources and utilizing the word-embedding model, the topic model, and deep transfer learning. The process involves identifying technology fields using sentence vectors that incorporate external semantic knowledge, which addresses the limitations of previous models that only consider word co-occurrence relationships. Real-time Google search data is also integrated to ensure the results are up-to-date. The proposed method was applied to photovoltaics technology and demonstrated impressive performance in enhancing topic coherence and predictive accuracy. The findings indicate that solar cell devices and materials, such as polymer materials and flexible photovoltaic devices, are the most promising technology opportunities based on multi-source data analysis.

Keywords Technology opportunity discovery · Deep transfer learning · Generative pretrained transformer · Technology field identification · Link prediction · Photovoltaics technology

Introduction

Technological innovation is the inexhaustible power of a country's prosperity. For many enterprises, technological innovation is the lifeline for their survival and development (Kumar et al., 2024). To promote technological innovation, many countries have introduced a series of relevant policies. For example, in 2018, the National Science & Technology Council of America issued *the Strategy for American Leadership in Advanced Manufacturing* (NSTC, 2018). The report highlighted the necessity of developing and advancing new manufacturing technologies, such as high-performance materials. Similarly, the Solar

✉ Lu An
anlu97@163.com

[1] Center for Studies of Information Resources, Wuhan University, Wuhan 430072, China

[2] School of Information Management, Wuhan University, Wuhan 430072, China

[3] Institute of Data Intelligence, Wuhan University, Wuhan 430072, China

Power Development *"Twelfth Five-Year" Plan* of China (MOST, 2012) has played a crucial role in guiding the development of photovoltaic (PV) technology.

Technology opportunity discovery (TOD) refers to the discovery and identification of potential opportunities for innovation and development in technology fields (TFs) (Klevorick et al., 1995; Li et al., 2019). Based on TOD, countries and enterprises can gain competitive advantages in identifying emerging technologies and research and development (R&D) projects (Al-Emran & Griffy-Brown, 2023). This enables them to focus their resources on TFs, seize development opportunities, and gain a favorable position in future technological competition.

Existing TOD methods predominantly rely on patent and academic literature analyses, employing techniques such as citation network analysis (Ampornphan & Tongngam, 2020) and topic modeling (Ma et al., 2022). While these approaches have demonstrated effectiveness in capturing historical technological trends, several inherent limitations remain. Traditional topic models, such as Latent Dirichlet Allocation (LDA), rely exclusively on word co-occurrence relationships, failing to capture deep semantic structures. More advanced deep learning models, such as Bidirectional Encoder Representations from Transformers (BERT), provide improved semantic representations but are constrained by input length limitations (Devlin et al., 2019). Furthermore, most existing TOD studies adopt a retrospective approach based on patent and literature data, which are subject to significant time lags due to lengthy publication time, limiting the capacity to detect real-time technological shifts.

To address these limitations, this study proposes an integrated research framework that combines generative neural networks, topic modeling, and deep transfer learning to enhance TOD. The Generative Pretrained Transformer—LDA—Autoencoder (GPT-LDA-AE) model is constructed to improve the accuracy of technology field identification (TFI) by incorporating contextual information into topic modeling, generating more coherent and meaningful topics while overcoming input length limitations. In addition, the substantial time lag associated with patents and literature due to extensive review and publication processes remains a critical challenge. To mitigate this issue, deep transfer learning is employed to integrate real-time web search data into TOD. Specifically, knowledge learned from labeled patent and literature data is transferred and adapted through adaptive transfer learning to unlabeled, real-time data sources such as Google search data. This approach enhances technology link prediction while ensuring the timely identification of opportunities within frontier TFs.

Related research

Technology field identification

Technology field (TF) is formed based on the classification of science and technology fields, combined with the characteristics of science and technology development, and serves to categorize some of the development priorities (Kang et al., 2023). Technology field identification (TFI) is an important method to clarify the status of technological development. It is also a prerequisite for the subsequent analysis of technological evolution and TOD. Patents, as one of the most important technological data (Kwon et al., 2022), are often considered good proxies for analyzing inter-technology relationships (Lee, 2021) and serve as a direct indicator of technological innovation, documenting novel inventions

and technical details. Meanwhile, academic literature provides theoretical foundations and insight into emerging research trends. Combining these two sources enables a comprehensive understanding of technological advancements, balancing practical applications with conceptual development. Thus, numerous studies have chosen to propose TFI methods using patents and literature as research data. Patent-based research methods focus on patent citation networks (Ampornphan & Tongngam, 2020). This is also the method used by some mainstream literature analysis software such as CiteSpace (Moral-Muñoz et al., 2020) and VOSviewer (van Eck & Waltman, 2010) to analyze indicators such as TFs and research hotspots. Several studies have utilized patent citation networks to identify TFs with high technological impact and development potential, to provide references for investment decisions and directions of technological development (Cammarano et al., 2023; Cho et al., 2021), to depict the dynamic patterns of knowledge change and structure among TFs (Andreoni et al., 2021; de Paulo et al., 2023; Jaffe & de Rassenfosse, 2017), and to assess and monitor the phenomenon of technological convergence (Chand Bhatt et al., 2021; Lee & Sohn, 2021; Son et al., 2020).

In addition to citation networks, some studies have used topic models to identify TFs (Ma et al., 2022; Song & Suh, 2019). For instance, Song and Suh (2019) utilized a latent Dirichlet allocation (LDA) topic model and network analysis tools to analyze patent information and identify major security fields. However, LDA only considers word-to-word co-occurrence information and cannot incorporate external semantic knowledge, leading to low-quality generated topics. In recent years, more advanced natural language processing techniques have been integrated into topic models (Li et al., 2019). For example, Talebpour et al. (2023) used the bidirectional encoder representation from transformers (BERT) (Devlin et al., 2019) to generate topics, and at least one variant of BERT outperforms LDA for any given metrics. However, BERT restricts the input to 512 characters, so that the sentence semantics may be partially lost due to the cropping of the input data in long text topic generation. It is evident that the topic model still needs refinement to balance the accuracy of topic generation and the semantic integrity of the input data.

Technology opportunity discovery

TOD is of great strategic significance for enterprise technological innovation and industrial development. According to previous research (Noh et al., 2016), TOD can be categorized into three aspects: future technology trends, technological innovation combinations (TICs), and key technologies. Among them, TICs refer to the combinations of technologies that have strong development potential in future but have not yet appeared (Gui & Xu, 2021).

Currently, there are two main approaches for TOD. The first is qualitative analysis, which relies on expert knowledge and specialized methods like the Delphi method and the analytical hierarchy process (Cho & Lee, 2013; Lee et al., 2014). Initially, companies relied on expert judgment to deal with the uncertainty of new technologies (Schaller et al., 2019). While experts are professional, they may not always be correct, especially as the amount of technical data increases (Schaller et al., 2019). Qualitative analysis may also be time-consuming and mixed with personal biases (Shen et al., 2020; Wang et al., 2021). Thus, some scholars have turned to quantitative analysis of objective data, including bibliometric and text mining techniques (da Silva Barboza et al., 2021; Jiang & Liu, 2023; Ren & Zhao, 2021). For example, Ma et al., 2022) used data mining, text mining, and terminology mining to identify technology opportunities (TOs) from patents.

Some studies have used link prediction, a network structure-based algorithm, for TOD based on supervised learning. The goal of link prediction is to predict whether two nodes are connected (Aaldering & Song, 2019; Gui & Xu, 2021). For instance, Yoon and Magee (2018) used keyword link prediction to initiate TOD for 3D printing, water purification, nuclear fusion, and other areas. The citation relationship in patent and literature data produces co-occurring networks that serve as labeled values required for link prediction in TOD (Wang et al., 2024). This is why researchers prefer patent data. However, the publication of patents and literature usually takes several months or more, resulting in a significant delay that does not respond well to the latest technological advances.

In recent years, the emergence of transfer learning has provided a powerful solution to many deep learning tasks that lack sufficient labeled data. Transfer learning is a technique that leverages knowledge from a source domain where labeled data is abundant to improve performance in a target domain where labeled data is scarce or unavailable. Due to its ability to handle data scarcity and domain adaptation challenges, transfer learning has been successfully applied in various fields. For instance, in the field of computer vision, transfer learning has been widely used to improve the performance of image classification models when labeled data is scarce (Kim et al., 2022; Yu et al., 2022a, 2022b). Mei et al. (2022) demonstrated that transfer learning can significantly enhance model accuracy by transferring knowledge from large-scale datasets (e.g., ImageNet) to small domain-specific datasets. Similarly, in natural language processing, transfer learning has been successfully applied to tasks such as sentiment analysis and text classification (Catelli et al., 2022). Adimulam et al. (2022) introduced a transfer learning framework for low-resource text classification, achieving state-of-the-art performance by fine-tuning pre-trained language models on small datasets. These studies highlight the versatility and effectiveness of transfer learning in handling data scarcity and domain adaptation challenges.

Given the success of transfer learning in other domains, it holds great potential for addressing the limitations of traditional TOD approaches, particularly the time lag of patent and literature publication. By leveraging the structured and labeled data from patents and literature (source domain), transfer learning can be adapted to analyze unlabeled, real-time data from Google search results (target domain). This approach ensures that the identified technology opportunities reflect the latest advancements, thereby providing a dynamic and responsive approach to technology opportunity discovery.

In summary, previous studies have either failed to accurately capture textual semantics due to limitations in text length for topic generation, or relied heavily on data such as patent and literature, which suffer from significant time lags. This study takes a step further by integrating the complete semantics of the data into the process of identifying technology domains and enhancing the timeliness of the findings through the inclusion of real-time data from Google search. Furthermore, by leveraging transfer learning—a technique proven effective in other domains, this study ensures that the identified technology opportunities reflect the latest advancements, offering a dynamic and responsive framework for technology opportunity discovery.

Theory of transfer of learning

The concept of transfer of learning first appeared in psychology and education (Royer, 1979). It refers to the process where experience and knowledge gained from previously learned related tasks positively influence the learning of a new task. Transfer learning is based on the theory of transfer of learning. Transfer learning aims to improve learning

performance in a target domain by utilizing knowledge from one or more source domains. Transfer learning can address issues such as data scarcity, cross-domain problems, and model training speed.

With the development of deep learning, deep transfer learning has gained research attention. The success of deep transfer learning is based on the powerful representation learning capabilities of deep networks (Zhuang et al., 2020). By using deep neural networks, deep transfer learning can better capture and represent complex features in data, leading to better performance across a wider range of tasks.

In transfer learning, the goal is for the algorithm to learn an optimal model in the target domain without labels, using the source domain. During this process, techniques are applied to minimize the data distribution differences between the source and target domains. In the context of deep learning, transfer learning can be represented as Eq. (1):

$$f^* = \arg\min \frac{1}{B} \sum_{i=1}^{B} \ell\left(f\left(x_i\right), y_i\right) + \lambda \mathfrak{R}\left(\mathrm{B}_s, \mathrm{B}_t\right) \tag{1}$$

where, B denotes the number of samples in a batch of data in deep learning, B_s and B_t represent a batch of samples from the source and target domains, respectively. f represents the objective function, $\ell(*, *)$ denotes the loss function. $\mathfrak{R}\left(\mathrm{B}_s, \mathrm{B}_t\right)$ is the regularization term, measuring complexity, and λ is the weight parameter balancing the two parts.

Unsupervised transfer in the target domain typically employs a dual-stream structure, consisting of a network with two branches. First, a batch of samples containing representations from both the source and target domains is input. Then, these samples pass through the shared weight layers of the first L layers (where the parameters of the neurons in both the source and target domains are completely shared) and enter the high-level feature layers. These high-level feature layers are the core of transfer learning, where most transfer operations occur. Finally, the network reaches the output layer, completing a forward pass.

During backpropagation, the network usually updates parameters using mini-batch SGD. Specifically, the network calculates the supervised learning loss (e.g., cross-entropy loss) using the true labels from the source domain and the predicted labels, and then uses the transfer loss computed by the bottleneck layer to jointly optimize Eq. (1). The bottleneck layer is typically a layer with fewer neurons than the input layer, allowing for more compact feature representation and significantly increasing training speed (Ganin & Lempitsky, 2015).

Methodology

Research framework

This research proposes a method for studying TOD using topic generation and deep neural networks. See Fig. 1. The method has three parts. First, a topic generation model GPT-LDA-AE is created by combining word-embedding technology, topic model, and generative neural network. This model identifies the main TFs of technology by integrating semantic relations. Second, the frontier fields of technology are obtained based on sequential topic association rules. Lastly, the technology map is used to explore gaps in the frontiers of the TFs. To eliminate low-relevance TICs in patent/literature data, a deep learning link prediction model is constructed. Deep transfer learning is then applied to TOD in

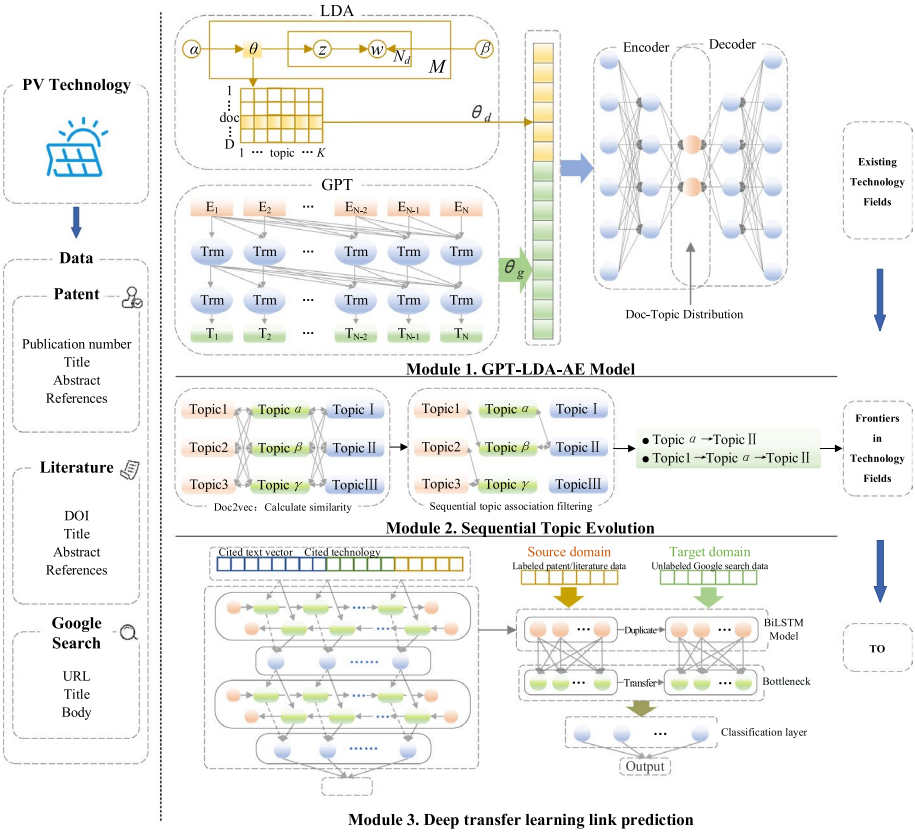

Fig. 1 Flowchart of the methodology in this study

Google search data. Finally, TOs with high development probability in multi-source data are obtained.

As shown in Fig. 1, research data includes patents, literature, and Google search data. The methodology consists of three major components. Module 1 performs technology field identification using the proposed GPT-LDA-AE model. Module 2 conducts topic evolution analysis to identify frontier technology fields. Module 3 focuses on technology opportunity discovery through BiLSTM-based deep transfer learning. Module 1 provides Module 2 with technology fields in different time windows. The identification of evolving frontier technology fields in Module 2 functions as a filtering and prioritization mechanism for Module 3. Specifically, only those TFs that exhibit nascent emergence and sustained evolutionary trajectories are used to construct the technology map and generate technology innovation combinations.

GPT-LDA-AE model and technology field topic generation

GPT is a pre-trained language model based on the structure of a transformer network, proposed by the OpenAI team (Radford et al., 2019). Compared to traditional methods like bag-of-words or TF-IDF, GPT can generate better word embeddings that capture

semantic information and can be used directly in downstream tasks. To generate the initial document-topic distribution, the study uses the LDA model, which is suitable for long-text topic generation. The input document is embedded into the vector space using the GPT word-embedding model, which captures the contextual content of the document and generates sentence vectors with contextual semantic information. This optimizes the topic generation results of the LDA model, which lacks semantics. The vectors generated by both models are located in a high-dimensional space with sparse information. Thus, the study introduces AE to learn the low-dimensional representation of high-dimensional vectors. The structure of the specific model is shown in Fig. 1.

The LDA model uses Gibbs sampling to obtain the topic distribution θ_d (K-dimensional vector) of the document d. The GPT model is trained on the document d to generate a high-dimensional vector θ_g (with dimension $G=768$) containing overall information such as sentence position, order, meaning, etc. The input vector v_a of the AE is formed by connecting the initial document-topic distribution θ_d and the high-dimensional sentence vector θ_g. The Encoder learns the mapping function from v_a to the latent representation θ_t (K-dimensional document-topic distribution) in the low-dimensional space through a 3-layer Dense network. In contrast to Encoder, Decoder constructs a mapping function from a K-dimensional document-topic distribution to the union vector v_a mapping, with v_d as the output. Specifically, the first layer of the Encoder is used as input for the high-dimensional vector v_a. The second layer of the network normalizes the input vectors and uses ReLU as the activation function. The third layer obtains the final output vector θ_t of Encoder by means of the Softmax activation function. θ_t is the output of GPT-LDA-AE model, which represents the document-topic distribution corresponding to document d, i.e., the probability that d belongs to each topic.

During training, GPT-LDA-AE model aims to minimize the error between input and output vectors in the AE. Additionally, it must maintain a high level of coherence between topics in the generated topic-word distribution. Once the training is finished, the model can map documents to topics and generate the word distribution for each topic. The mean absolute error (MAE) is used to measure the distance between the input v_a and the output v_d of the AE, as shown in Eq. (2).

$$Loss = \left| v_a - v_d \right| \tag{2}$$

The training process of GPT-LDA-AE model is as follows.

Step 1: Initialize the weights W and the bias b of AE, the maximum number of topics K and other related parameters.
Step 2: Initialize the number of topics $k=1$.
Step 3: Input the preprocessed documents into GPT and LDA, respectively, to obtain the output vectors θ_g and θ_d. Concatenate θ_g and θ_d to obtain the joint vector v_a and input it into AE.
Step 4: Obtain the output error of AE according to Eq. (2). Optimize the weights and bias terms of each layer of the network of AE by the back propagation algorithm until the model converges.
Step 5: The output θ_t of Encoder, i.e., the document-topic distribution, is obtained from the converged AE, and the word distribution v_p for each topic is further obtained according to Eqs. (3)-(4).

$$p_k^w = \sum_{d=1}^{D} N_{dw} p_k \Big/ \sum_{w=1}^{V} \sum_{d=1}^{D} N_{dw} p_k \tag{3}$$

$$v_p = \left\{ p_k^1, p_k^2, \ldots, p_k^V \right\} \tag{4}$$

where V denotes the total number of words in the word set, which is obtained from each document after word segmentation. D is the number of documents. N_{dw} is the number of words contained in document d, and p_k is the element in the document-topic distribution, i.e., $\theta_t = \left\{ p_1, \ldots, p_k, \ldots, p_K \right\}$. K is the number of topics.

Step 6: The quality of generated topics is measured by the widely used *CV* consistency metric (Röder et al., 2015). Call the open-source library *gensim* to calculate and record the consistency score *c_v*.
Step 7: If $k \leq K$, k=$k+1$, go to Step3; otherwise, go to Step8.
Step 8: Determine the number of topics k that corresponds to the maximum value of *c_v*, run Step 4 and Step 5 to get the final topic-word distribution and the document-topic distribution.

Topic evolution and frontiers identification in technology fields

To determine whether there is a correlation between different topics in neighboring time windows, the cosine similarity is used as a measure. The degree of similarity between technological topics in sequential windows $\mathrm{CS}\left(T_i^t, T_j^{t+1}\right)$ can be obtained according to Eq. (5).

$$\mathrm{CS}\left(T_i^t, T_j^{t+1}\right) = \frac{\sum_{w=1}^{V} p_w\left(T_i^t\right) \times p_w\left(T_j^{t+1}\right)}{\left(\sum_{w=1}^{V} \left(p_w\left(T_i^t\right)\right)^2\right)^{1/2} \times \left(\sum_{w=1}^{V} \left(p_w\left(T_j^{t+1}\right)\right)^2\right)^{1/2}} \tag{5}$$

where T_i^t is the i-th topic in the time window t. $p_w\left(T_i^t\right)$ and $p_w\left(T_j^{t+1}\right)$ are the probability of word w in the topic-word distribution in the time window t and $t+1$, respectively.

According to Dai et al. (2022), there are two directional correlations between topics in sequential windows, i.e., the forward topic and the backward topic. The forward topic of T_j^{t+1} can be defined as follows: Calculate the cosine similarity of each technical topic between the previous time window of T_j^{t+1}, i.e., time window t, and T_j^{t+1}. If there is T_i^t such that $\mathrm{CS}\left(T_i^t, T_j^{t+1}\right) \geq \mathrm{CS}\left(T_{k^t}^t, T_j^{t+1}\right)$ (where $k^t \in \{1, 2, \ldots, K^t\}$, K^t is the total number of technological topics in time window t), T_i^t is considered to be the forward topic of T_j^{t+1}, denoted as $For\left(T_j^{t+1}\right) = T_i^t$. Accordingly, the backward topic of T_i^t is denoted as $Back\left(T_i^t\right) = T_j^{t+1}$. If T_i^t and T_j^{t+1} are forward and backward topics of each other, they are considered to have an association relationship.

The associations between different sequential topics are not always valid. This section improves the association filtering rule between topics proposed by Wan et al. (2022) to improve the reliability of associations between sequential topics. The specific rules are as follows.

1) Similar to the citation (Wan et al., 2022), a minimum threshold for the existence of association relationships between topics is set: The average similarity between topics that have already established an association relationship is used as the threshold, and association relationships below the threshold are deleted;

2) To ensure the closeness and strength of association between two technological topics that have an associative relationship, the following rule is further set. When T_j^{t+1} is the backward topic of T_i^t, if one of the following conditions is satisfied, the association between T_j^{t+1} and T_i^t is considered invalid: all the values of $\mathrm{CS}\left(T_{k'}^t, T_j^{t+1}\right)$ are sorted in descending order, and the association between T_j^{t+1} and T_i^t is considered invalid if $\mathrm{CS}\left(T_i^t, T_j^{t+1}\right)$ is after the 4th position; if $\mathrm{CS}\left(T_i^t, T_j^{t+1}\right)$ is ranked between the 2nd and 4th position while $\mathrm{CS}\left(T_m^t, T_j^{t+1}\right) > \mathrm{CS}\left(T_i^t, T_j^{t+1}\right)$ exists, i.e., the position of T_m^t is in front of T_i^t and the backward topic of T_m^t is not T_j^{t+1}, then the association between T_j^{t+1} and T_i^t is invalid. When T_i^t is the forward topic of T_j^{t+1}, the discrimination of association validity is the same as in the previous section.

According to the knowledge evolution theory and the life cycle theory, the evolutionary relationship of topics in the technology field can be divided into five types, i.e., nascent, inherited, fusion, split, and extinction. The nascent type indicates that the current topic is newly created and did not exist in the previous time window. The inherited type suggests that the current topic is a continuation of the topic in the previous time window. The fusion type denotes those two or more topics from the previous time window are merged into a new topic in the current time window. The split type indicates that a topic from the previous time window is divided in the current time window to form two or more new topics. The extinction type signifies that a topic from the previous time window disappears in the current time window. In this study, based on the evolutionary relationship of sequential topics and previous research (Jin et al., 2020; Yang et al., 2022), the technological topics with nascent and capable of continuous evolution (including inheritance, fusion, or splitting) are considered to be the frontier field of technology. The identification of frontier technology fields not only reveals promising directions of technological advancement, but also serves as the analytical foundation for subsequent opportunity discovery. In the next section, these frontiers are further examined using deep transfer learning to uncover high-potential technology opportunities based on their linkage with current technological developments.

Technology opportunity discovery based on deep transfer learning

Technological mapping

Technology maps are an effective tool to discover current technology gaps and further identify TOs. In this section, generative topographic map (GTM) (Wang et al., 2022) is introduced to perform technology mapping on the frontiers in TFs obtained in the previous section. Thus, the current technology gaps are mined. The frontier TFs identified through topic evolution serve as the basis for the GTM process, where keywords of these topics are used to construct technology combinations, enabling the mapping and identification of innovation gaps within the evolving technological landscape. Based on the inverse mapping capability of GTM, the TICs indicated by each gap are discovered as shown in Fig. 2. First, high-frequency keywords are extracted from the frontiers in TFs to form the keyword

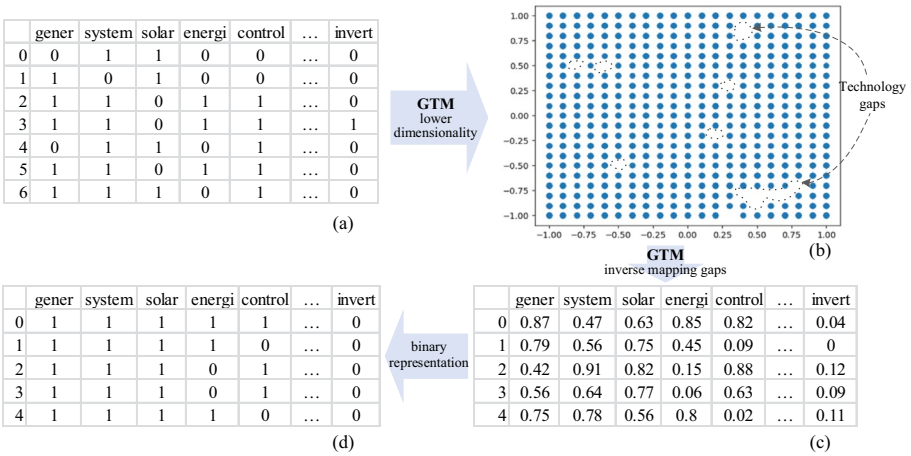

Fig. 2 The GTM process

0–1 vector of each document, which is the technology combination of each document (Fig. 2(a)). Then, the keyword matrix is run through the GTM algorithm to obtain the technology map (Fig. 2(b)) and the technology gaps. Finally, the technology gaps are mapped by the inverse mapping of GTM to obtain the TICs that they characterize (Fig. 2d).

BiLSTM link prediction model

Not all technology gaps are TOs for further development. Some gaps may be difficult to develop due to their lack of relevance to current technologies. However, if a technology gap is highly relevant to current technologies, it is likely to be developed in the future. This suggests that the development of new technologies comes from inheriting, fusing, and continuing existing technologies. To determine the probability of a technology gap being a TO for development, a BiLSTM deep learning link prediction model is proposed. This model performs link prediction between TICs and existing technologies to identify the most probable opportunities.

The presence of citations in patents and literature is evidence of technologies' continuous development over time. Hence, labels for link prediction datasets can be constructed based on the citation relationship. Table 1 shows the specific structure of the model input data. The technological combination of citing and cited documents comprises

Table 1 Patent/literature documents for link prediction

Citing Document	Cited Document	Technological Combinations		Link
		In the citing document	In the cited document	
CN105144403B	CN201360011Y	(solar,…, cell)	(batteri,…,light)	1
US9511877B2	KR2011047713A	(electr,…,surface)	(electrod,…,gener)	0
…	…	…	…	…
10.1021/ja906621a	10.1039/c3ee41272e	(wire, …, cover)	(devic, …, oxid)	0

high-frequency keywords stem extracted from each document. If there is a citation relationship between the documents, the link is labeled as 1.

The BiLSTM link prediction model framework is shown in Fig. 1 and the specific parameters of the model are shown in Table 2

In addition to the technological combinations in Table 1, the BiLSTM model also takes the text of the cited documents as input at the same time. In this way, the semantic information of the documents is obtained. The GPT model is used to generate sentence vectors of the text of the cited documents. These sentence vectors are combined with the keyword vectors of the technological combination, forming the input vectors for the BiLSTM link prediction model. The model is trained to take the link existence probability as the output to realize the probability prediction of TICs.

Deep transfer learning link prediction

Deep transfer learning can solve the problem of insufficient labeled data in the target domain through the knowledge learned in the source domain. This technique can also be used for unlabeled data by transferring the model from the source domain to the target domain and measuring the transfer loss between them. This study primarily utilizes patents and academic literature as source domains, given their complementary strengths in representing technological advancements. Additionally, the citation networks inherent in these data sources provide essential labeled relationships for training the link prediction model used in deep transfer learning. These labeled links serve as the foundation for adapting the model to unlabeled Google search data, enabling the identification of emerging technology opportunities from real-time sources. Alternative data sources, such as industry reports, generally lack such structured citation relationships, making them less suitable for the link prediction framework adopted in this study.

The form of Google search data is comparable to that of patents and literature, but there are slight variations in the text content. Simply applying the BiLSTM link prediction model to forecast TO in unlabeled Google search results would result in less reliable and less precise predictions. Thus, this study adopts the structure-adaptive deep transfer learning method based on the BiLSTM model described in the section of BiLSTM link prediction model. By learning the differences among the patents, literature and Google search results, the BiLSTM model can transfer from aiming at patents and literature (the source domain)

Table 2 Parameters of the BiLSTM link prediction model

Layer	BiLSTM1	units = 64, kernel_regularizer = L2(0.001), activation = tanh
	Dropout1	rate = 0.5
	BiLSTM2	units = 32, kernel_regularizer = L2(0.001), activation = tanh
	Dropout2	rate = 0.5
	Dense	units = 1, activation = sigmoid
Compile	loss = binary_crossentropy, optimizer = adam, metrics = accuracy	
Fit	epochs = 50, batch_size = 64	

to being used for unlabeled Google search results (the target domain). The specific structure of the model (See Table 3) is similar to the description in the section of BiLSTM link prediction model. In addition, to bridge the gap between the source (patent/literature) and target (Google search) domains, the transfer learning model introduces a bottleneck layer for feature alignment and includes both classification and transfer loss in the training objective. This allows the model to generalize the citation-based link prediction capability from labeled data to the unlabeled real-time domain, enabling dynamic discovery of emerging technology opportunities.

Based on the initial parameters, the model uses the knowledge learned in the source domain to adjust the model structure accordingly, thus improving its performance in the target domain. By converging models, link prediction between unlabeled Google search results and TICs can be accomplished.

Results and discussion

Data collection and processing

Solar energy is a promising renewable energy source (Han et al., 2022). Despite strong policy support (DOE, 2016; MOST, 2012), the momentum of innovative development in photovoltaic (PV) technology has weakened. It is urgent to use available data to identify potential development directions for PV technology, and guide technicians to accelerate the R&D. This study uses three types of data, i.e., patents, literature, and Google search. Patents reflect technological innovations and development trajectories. Literature provides theoretical foundations and emerging research insights. Google search data captures real-time industry trends and public discourse. The combination of these sources ensures both long-term technological depth and up-to-date relevance. More importantly, both patents and literature contain structured citation networks, which can be leveraged as labeled data for training deep learning models in link prediction tasks during technology opportunity discovery. These citation relationships allow us to construct a supervised learning dataset for deep transfer learning, which is essential for adapting the model to unlabeled Google search data. Without such labeled relationships, it would be challenging to effectively transfer knowledge from structured datasets to real-time web-based information sources.

To obtain patents, the Derwent Innovations Index was searched using the search query "((TS = (PV power generation)) OR TS = (photovoltaic power generation)) OR

Table 3 Parameters of the deep transfer learning model

Structure	Base network	BiLSTM	units = 64, activation = tanh
		Dropout	rate = 0.5
		BiLSTM	units = 32, activation = tanh
		Dropout	rate = 0.5
	Bottleneck	Dense	units = 128, activation = relu
	Classification	Dense	units = 2, activation = softmax
Compile	loss = crossentropy, transfer loss = mse, optimizer = adam, lr = 0.003, lr_gamma = 0.0003, lr_decay = 0.75		
Fit	epochs = 50, batch_size = 64		

TS = (photovoltaics)". A total of 44,401 patents related to the PV technology were published from 1938 to the time of the search (September 1, 2022). Information such as the publication number, title, abstract, publication year, and cited patents were selected and exported as research data.

Literature was obtained from the Web of Science. A total of 25,939 articles published since 1984 were retrieved from this database using the same query. The digital object unique identifier (DOI), title, abstract, year of publication, references, and other information were selected and exported.

Google search results were obtained by a Python crawler. Since only 6 patents existed before 1970 and no literature was available, only Google search results from 1970 to the present were crawled. The crawled information included the uniform resource locator (URL), publication time, title, and body content. A number of 12,305 relevant search results was obtained.

The data was subjected to necessary cleaning operations before it was used: 1) deleting data missing publication numbers, DOIs, or URLs; and 2) deleting data where both titles and abstracts were missing.

Both TFI and TOD require accurate word-level input data. Thus, the data underwent word segmentation which included the following steps: 1) unifying the form and converting all words to lowercase; 2) removing punctuation, numbers, and consecutive repeated words and phrases; 3) filtering meaningless text such as stop words; 4) extracting common and proper nouns, with proper nouns extracted by constructing a proper noun dictionary as an aid; and 5) extracting stems. After processing, a number of 79,297 usable pieces of data were obtained, including 41,404 patents, 25,596 literatures, and 12,297 Google search results.

This study used a pre-discretization approach to analyze the data. Firstly, all the data was classified into different time windows based on the year of publication. Then, the data in each time window were analyzed separately. Due to being too scattered and relatively old, the pre-2000 data was analyzed as a collection. Additionally, the post-2000 data was divided into five time windows, i.e., 2001–2005, 2006–2010, 2011–2015, 2016–2019, and 2020–2022. As a result, a total of six datasets were obtained, and each dataset contained 7,469, 3,242, 7,808, 21,951, 15,515, and 23,312 pieces of data, respectively.

Existing technology field identification of PV

Comparison of model performance

To assess the effectiveness of GPT-LDA-AE model for topic modeling, its performance is compared with the classical LDA topic model, the neural variational document model (NVDM) (Miao et al., 2015), and the adversarial-neural topic model (ATM) (Wang et al., 2019). Five different topic settings ([5, 10, 15, 20, 30]) are applied to the PV technology dataset. The average topic coherence is calculated through multiple experiments. Moreover, to ensure comprehensive and accurate topic coherence, CA (Aletras & Stevenson, 2013), NPMI (Aletras & Stevenson, 2013), and UCI (Newman et al., 2010) are added to the CV indicator to measure topic coherence. See Fig. 3. A higher score for all the three types of indicators indicates higher topic coherence.

As shown in Fig. 3, GPT-LDA-AE model outperforms other topic models in different number of topics and different coherence metrics. The results show that the introduction of contextual semantic relationships and the strong characterization ability of neural networks

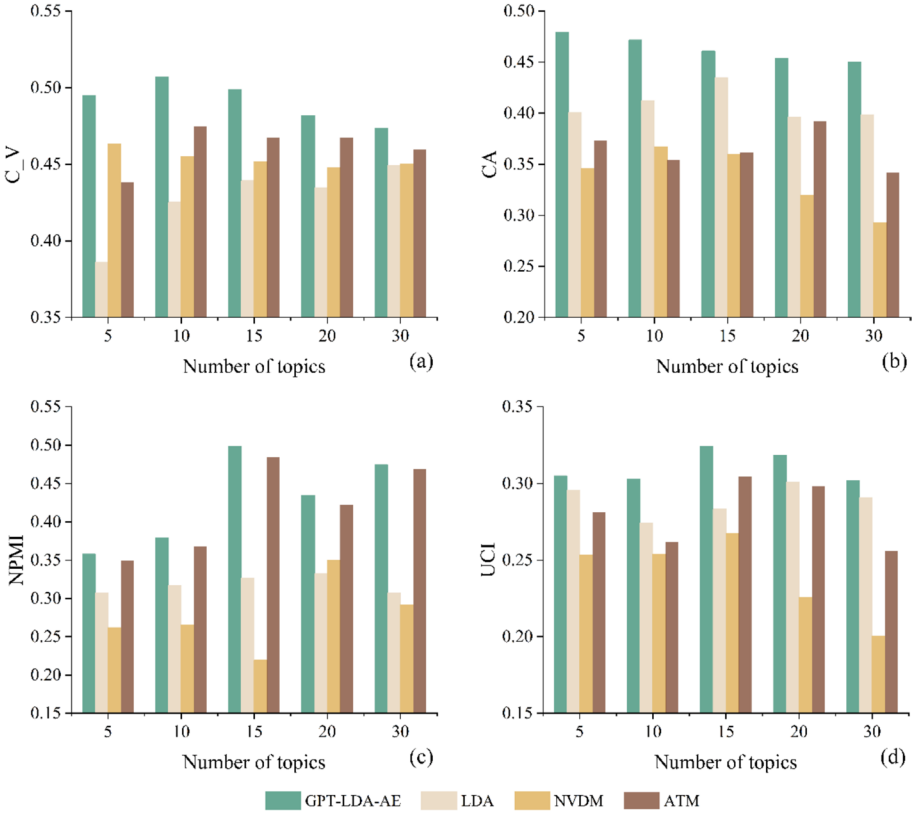

Fig. 3 The number of topics and the coherence score of GPT-LDA-AE and comparison models

in topic generation enable the model to fit the real data distribution accurately and thus to mine high quality topics.

Results of PV technology field identification

To avoid too scattered topic content, set $K=30$. According to the change of *CV* value with the increase of the number of topics, the optimal number of topics for the sequential window is set to 10, 10, 17, 14, 14, and 11 in order.

The analysis by GPT-LDA-AE revealed that there were ten major TFs in the global PV industry during its initial development phase before 2000. This phase was a long period that began with the publication of the first patent related to the PV technology. The patent was for a solar power generating device, and it corresponded to Topic 5. It mainly covered the field of solar thermal power generation technology, which included technologies such as water distillation, solar pumping, and low-boiling fluid circulation power generation. The other TFs identified focused on technologies such as solar condensers and collectors, solar electrolysis devices, PV power conditioners, solar cell arrays, compound semiconductor solar cells, user-side solar power generation, distributed PV transmission, and solar vehicles.

The period of 2001–2005 primarily focused on battery materials, solar drives, PV system control, PV system development, and cost analysis, along with solar thermal utilization. The following period of 2006–2010 brought about new TFs related to transparent conductive oxide films, polymer fullerene solar cells, silicon deposition technology, solar cell substrate materials, and nanomaterials. The years 2011–2015 saw a rapid development of PV technology, with significant growth in relevant patents, literature, and Internet discussions. Compared with the previous periods, greater progress was made in the fields of organic solar cells, thin-film solar cells, and concentrating PV, further improving PV efficiency and lowering the cost of PV commercialization. In the period of 2016–2019, the growth rate of technology research slowed down.

In the latest time window, scientists and technologists mainly conducted research in 11 TFs, including system control and power output efficiency, renewable energy technology development, solar panel support structures, PV system monitoring/fault detection, energy storage, optimal configuration of PV system, new solar cells, photothermal conversion, system efficiency, and output technologies, PV system evaluation methods, and solar cell materials.

PV technology evolution and frontier field identification

The valid topic associations among the topics in sequential windows are obtained after topic correction and filtering. There are 6, 8, 12, 11, and 12 pairs, respectively. The visualization of each topic correlation in sequential windows showed the evolution relationship between topics in different time windows. See Fig. 4, where 2005#8 represents the 8th topic in the time window between 2001 and 2005.

Figure 4 shows that there are 4 paths in the latest 4 time windows that are nascent and have continuous evolution relationships. The 4 evolutionary paths contain a total of 14 relevant topic nodes, i.e., the PV technology frontiers, and their corresponding specific topic terms are described in Fig. 5.

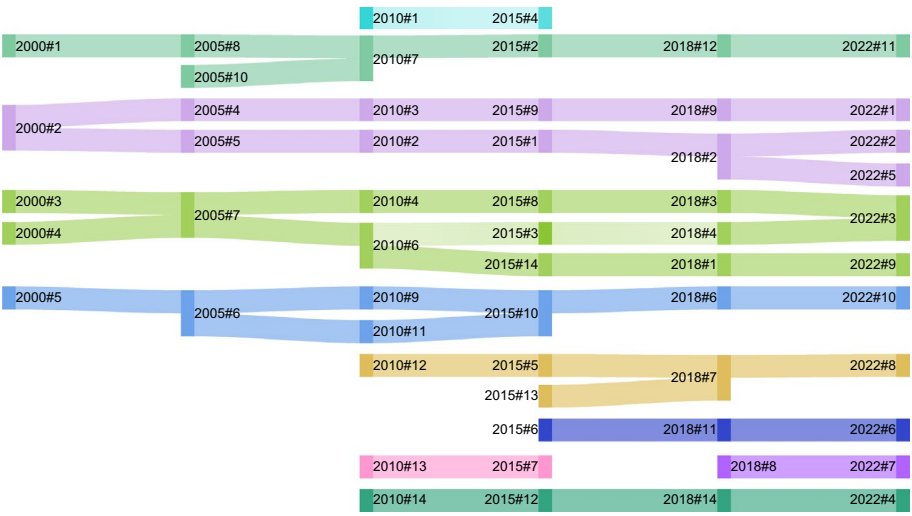

Fig. 4 Sequential topic evolution relationships

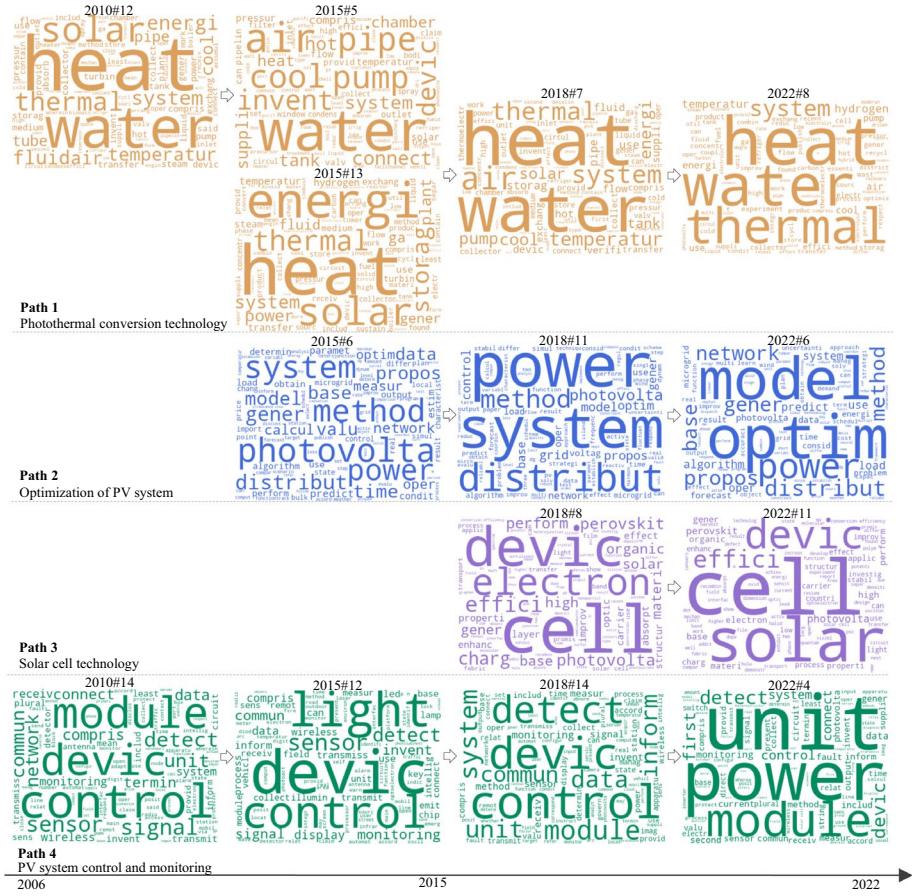

Fig. 5 Frontiers in technology fields

The first evolutionary path mainly involves the innovative development of photothermal conversion technologies and their applications, especially the innovation of combined solar and thermal power (CST). Initially, CST systems used tower reflectors to concentrate the sun rays onto collectors. Later, the CST system was improved and the types of collectors became more diverse (2010#12, 2015#13, 2022#8, etc.). These include tank-type collectors, tower-type collectors, and so on. During the development of CST, key technologies have been further enhanced, such as improvements in optical performance, heat transfer efficiency (2022#8), and durability of collectors, as well as the development of energy storage technology (2018#7).

Path 2 emphasizes the importance of PV system modeling and optimization. Path 3 is all about innovation and development of solar cell technology, with a special focus on improving PV conversion efficiency. Stakeholders have been directing their efforts towards improving this efficiency. In recent times, the focus has shifted from carrier transport and recombination problems to the stability and reliability of PV technology in practical applications. This has been possible due to the emergence and development of new solar cells such as perovskite solar cells, organic solar cells, quantum dot solar

cells, etc. Path 4 involves the control and monitoring of PV systems, including data processing, communication, and image recognition technologies.

PV technology opportunity discovery

Map of PV technology

High-frequency keywords in the frontiers of TFs were extracted using Python. The GTM model of *the ugtm* open-source library was used to draw technology maps of patents, literature, and Google search results related to the PV technology. The model selected the 41*41-dimensional Gaussian function as the base function. The posterior plurality mode projection of the data was used to generate the technology maps. See Fig. 6.

As shown in Fig. 6, the technology combinations in the current patents mainly concentrate on the left side of the map. The technology combinations in the literature mainly concentrate on the right side and the lower part of the map. This suggests that there is a difference between the fields where the PV technology is currently under scientific research and those that have entered the experimental development stage. Compared to patents and literature, the technology combinations in the current Google search results are relatively small, and there are more potential TICs.

Using the GTM, the technology gaps in Fig. 6 are inversely mapped to the probability of occurrence of each technology keyword in the original space. According to Gui and Xu (2021), the probability threshold is set to 0.3. We filter to obtain a 0–1 matrix of high probability technology keywords, and each row of the matrix corresponds to a set of potential TICs. Finally, 165 TOs were obtained.

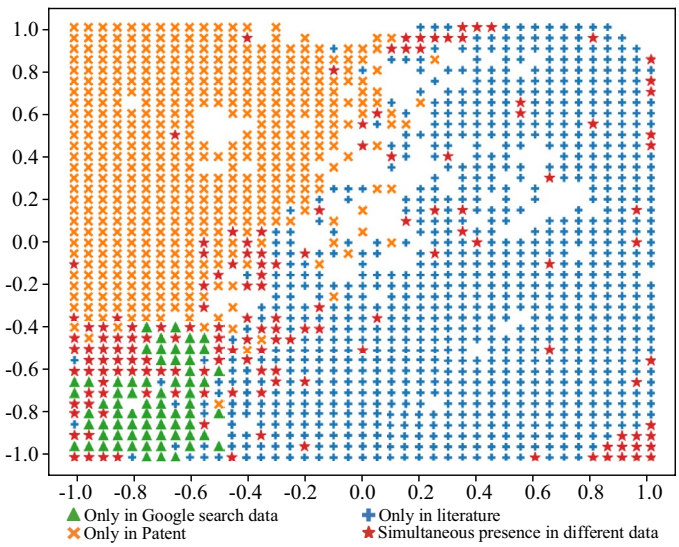

Fig. 6 Map of photovoltaics technologies based on multi-source data

PV technology opportunity discovery based on deep transfer learning

The investigated patents and literature and their citation data are used to construct the dataset for model training. The training set, validation set and test set are divided by 80%, 10% and 10%, respectively. To avoid temporal information leakage when using citation relationships as labels, the training, validation, and test sets were divided chronologically based on the publication year of the documents. Specifically, all documents in the validation and test sets were published after those in the training set, ensuring that the model only learns from past data to predict future citation links. This temporal split more accurately reflects the real-world dynamics of technological development and prevents data contamination. This study is based on Python 3.7 and Keras 2.3.1, and a computer with a CPU i9-9960X is used for calculation. After training, the BiLSTM link prediction model reaches convergence at epoch = 36 with loss = 0.001, which provides good prediction results. To verify the prediction performance of the model, the BiLSTM model is compared with the newer prediction model AMPSO-CLSTM (max_filters: 512, kernel_size: 2, stride: 1, max_hidden_units: 100) (Yu et al., 2022a, 2022b). Meanwhile, commonly used models, including random forest (RF), decision tree (DT), and support vector machine (SVM), are compared. The recall, precision, and F-score are used as performance metrics and the results are shown in Table 4.

Table 4 shows that the BiLSTM link prediction model proposed in this study has high prediction performance, indicating that the model can be effectively applied to the link prediction of patents and literature. It is applied to the link prediction between the current patents/literature and the TICs in Section Map of PV technology. Eleven TICs with a link existence probability higher than 0.9 are obtained. See Table 5.

Each TO contains several technology keywords. Based on the keywords, R&D personnel can determine the future direction of a particular field of the PV technology and thus realize technological innovation. For example, the keyword stems in TO 6 includes *photovolta*, *plural*, *flexibl*, *organic*, *polym*, *nanoparticl*, *metal*, *coat*, *substrat* and so on. This TIC is related to the design of flexible substrates in organic PV. Technicians can embark on innovative research on existing flexible substrates from the perspective of metal nanoparticles and polymers.

According to related research (Allouhi et al., 2023; Khan et al., 2022; Schmid, 2023), the future innovation direction of the PV technology mainly focuses on the transformation of polysilicon preparation process route, PV grid-connected technology, thin-film materials, PV core devices with high efficiency and low-cost low-energy development, and other TFs. Among them, high photoelectric conversion efficiency, high power, and stable output cell technology will be the focus of future innovation research. This is highly relevant to TOs 1, 6, 8, 9, and 11 in Table 5, which include the devices and materials of solar cell such as polymer materials, flexible PV devices, perovskite solar cells, flexible materials, and

Table 4 Comparison of model prediction performance

Model	Recall	Precision	F-score
BiLSTM link prediction model (this study)	**0.904**	**0.954**	**0.956**
AMPSO-CLSTM	0.872	0.922	0.914
RF	0.853	0.901	0.906
DT	0.803	0.875	0.854
SVM	0.837	0.892	0.899

Bold values highlight the top-performing entries across each column

Table 5 PV technology opportunities in patents and literature

No	TO (Keyword stem)	Probability
1	cell, pv, electr, oper, heterostructur, solar, field, energy, electron, light, engin, convert, layer, surfac, semiconductor, manufactur, deposit	0.986
2	electroluminesc, thermal, imag, light, soak, pid, spectral response, defect, map, hot, spot, iv, curv, quantum efficiency, reliabl	0.977
3	unit, power, detect, control, connect, system, monitor, ingener, output, invent, suppli, plural, electr, calcul, inverter, remote, diagnosi	0.969
4	control, system, output, oper, calcul, safeti, perfunction, distribut, algorithm, load, uncertainty, variabl, integr, coordin, effici, phase, limit	0.964
5	microgrid, load, model, identifi, parallel, compon, aerial, spot, junction, disclos, diod, smart, solar, strategi, neural, reliabl, evalu	0.957
6	cell, plural, process, multi, cost, reduc, flexibl, organic, applic, light, polym, nanoparticl, layer, surfac, metal, coat, substrat	0.937
7	microgrid, smart, grid, energi, storag, demand, respons, energi, manag, system, system, optim, grid, tie, island, ac, coupl	0.925
8	temperatur, compon, energi, materi, applic, structur, polym, transfer, fabric, mechan, dimension, molecular, liquid, composit, galss, poli, fiber	0.911
9	power, devic, flexibl, effici, perovskit, optic, fabric, film, quantum, dot, bandgap, pce, layer, mental, electrod, transpar, photon	0.907
10	control, system, photovolta, gener, output, electr, temperatur, cell, plateffect, hybrid, materi, phase, concentr, thermal, cool, thermoelectr, coupl	0.904
11	photovolta, convers, cell, perreduc, evalu, polym, fabric, fulleren, heat, water, thermal, rate, evaporation, surfac, transpar, insul, transmitt	0.901

thin film technologies, bandgap modulation technologies, and transparent conductive materials and devices. In addition, TOs 2, 3, 4, 5, and 7 in Table 5 address PV system technologies. These include the optimal design of PV system operation and maintenance, monitoring, intelligent control and testing algorithms, and microgrid optimization. The goal is to improve power generation efficiency, power output reliability, and stability. TO 10 involves thermoelectric coupling technology. The comparison results also show that the BiLSTM link prediction model is able to effectively filter the technology gaps with low relevance, thereby obtaining TOs with high development probability, which can be applied to the prediction research of TOs.

In order to predict links between unlabeled Google search results and technology gaps, an adaptive deep transfer learning approach is used to retrain the BiLSTM link prediction model. The model reaches convergence at epoch 45, and its predictive performance is comparable to the original BiLSTM model, with accuracy stabilized at 0.952. The model is used for link prediction between the current Google search results and the TICs in Section Map of PV technology. Five TICs with a link existence probability higher than 0.9 are obtained. See Table 6, three of which are the same as those in the patents/literature.

As shown in Table 6, similar to the TOs in patents/literature, there are also TOs in Google search results for solar cell materials (TO1, 3 and 4), PV power system monitoring, diagnosis and remote control (TO5), and thermoelectric coupling (TO2), indicating that they have a high probability of becoming technological breakthroughs in the future. Experts should focus their research on these 3 directions. In addition, unlike TOs present in

patents and literature, TO3 and TO5 in Google search results emphasize different aspects. TO3 focuses on devices such as heterojunctions, quantum dots, graphene, and so on. TO5 focuses on identifying and resolving possible failures and defects, and optimizing the operation of the system by real-time monitoring and controlling using remote technologies.

To further validate the effectiveness of transfer learning in TOD, this study conducts an ablation study by removing the transfer learning component and assessing the model's performance when trained solely on the source domain data (patents and literature) and applied directly to the target domain (Google search data). Since the target domain lacks labeled data, direct supervised training is infeasible. Instead, an experiment is designed to compare two settings: 1) With transfer learning (original model): The BiLSTM link prediction model is first trained on structured patent and literature data, then adapted to Google search data using transfer learning techniques, including domain adaptation mechanisms. 2) Without transfer learning (baseline model): The BiLSTM link prediction model is trained exclusively on patent and literature data without applying any transfer learning mechanisms. The trained model is then directly used to predict technology opportunities in Google search data.

Given the absence of labels in the target domain, this study applies t-SNE to visualize how the feature representations of Google search data change before and after applying transfer learning. A well-structured embedding space suggests that transfer learning enhances domain adaptation. Figure 7 presents the t-SNE visualizations of the feature space for models with and without transfer learning. Due to the large dataset size, only a randomly selected subset of 3,000 data points is used for visualization.

The visualization shows that the target domain data become more structured and shift closer to the source domain with transfer learning. The results indicate that the transfer learning model produces well-separated feature representations, suggesting better adaptation of Google search data. In contrast, the baseline model without transfer learning results in a more scattered distribution, confirming that the model struggles to extract meaningful patterns from unstructured data without domain adaptation.

Conclusion

This study proposes a framework for technology opportunity discovery that integrates topic modeling and deep transfer learning. By leveraging the GPT-LDA-AE model for technology field identification and a BiLSTM-based link prediction model for technology opportunity discovery, the proposed method addresses key limitations of existing approaches, including reliance on static patent and literature data, time lag issues, and inadequate semantic representation in topic modeling. Compared to conventional methods, this approach enables a dynamic and accurate identification of emerging technological opportunities.

Empirical validation in the field of photovoltaics demonstrates the effectiveness of the proposed framework. The GPT-LDA-AE model outperforms baseline models in all topic coherence indicators, ensuring improved accuracy in technology field identification. Additionally, the BiLSTM model achieves a precision of 95.4% in link prediction. With deep transfer learning, the accuracy of the model remains stable at 95.2% when applied to unlabeled real-time Google search data, confirming its robustness in detecting high-potential opportunities. The case study reveals that technological advancements in photovoltaics have evolved significantly, with recent trends focusing on photothermal

Table 6 PV technology opportunities in Google search results

No	TO (Keyword stem)	Probability
1	temperatur, compon, energi, materi, applic, structur, polym, transfer, fabric, mechan, dimension, molecular, liquid, composit, galss, poli, fiber	0.956
2	control, system, photovolta, gener, output, electr, temperatur, cell, plateffect, hybrid, materi, phase, concentr, thermal, cool, thermoelectr, coupl	0.927
3	devic, control, photovolta, gener, circuit, heterostructur, junction, combin, electron, carrier, optic, band, film, quantum, optoelectron, dot, graphen, gate	0.915
4	cell, pv, electr, oper, heterostructur, solar, field, energy, electron, light, engin, convert, layer, surfac, semiconductor, manufactur, deposit	0.906
5	inorgan, exchang, defect, appli, comput, optim, identifi, acquisit, identif, meter, function, aerial, block, monitor, ground, diagnosi, remot	0.901

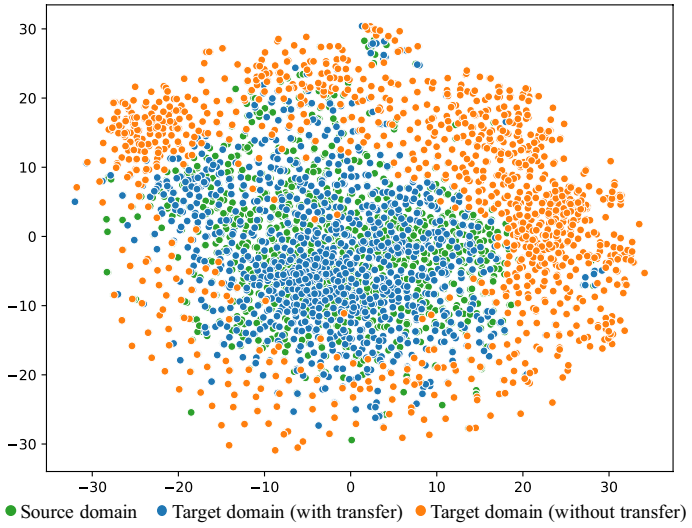

Fig. 7 Feature distributions before and after transfer learning

conversion, PV system optimization, and solar cell innovations. Further analysis highlights polymer materials, flexible PV devices, and thermoelectric coupling technology as key areas with breakthrough potential.

Beyond the photovoltaics sector, the proposed framework can be applied to various technology-intensive fields, providing a scalable and data-driven method for identifying emerging technology directions. By enhancing both the accuracy and timeliness of technology opportunity discovery, this study contributes to innovation management, enabling policymakers, researchers, and industry stakeholders to make more informed decisions in resource allocation and strategic planning.

Acknowledgements This work was supported by the National Social Science Foundation of China (Grant No. 23&ZD230).

Author contributions Ruilian Han: Conceptualization, Methodology, Formal analysis, Programming, Writing – original draft and review. Lu An: Formal analysis, Project administration. Wei Zhou: Writing – original draft and review. Gang Li: Formal analysis, Writing and review.

Funding National Social Science Fund of China,23&ZD230,Lu An

Declarations

Conflict of interest The author does not have any competing interests to declare.

Ethical approval I certify that this manuscript is original and has not been published and will not be submitted elsewhere for publication while being considered by *Scientometrics*. And the study is not split up into several parts to increase the number of submissions and submitted to various journals or to one journal over time. No data have been fabricated or manipulated (including images) to support your conclusions. No data, text, or theories by others are presented as if they were our own. The submission has been received explicitly from all co-authors. And authors whose names appear on the submission have contributed sufficiently to the scientific work and therefore share collective responsibility and accountability for the results.

References

Aaldering, L. J., & Song, C. H. (2019). Tracing the technological development trajectory in post-lithium-ion battery technologies: A patent-based approach. *Journal of Cleaner Production, 241*, 118343.

Adimulam, T., Chinta, S., & Pattanayak, S. K. (2022). Transfer learning in natural language processing: overcoming low-resource challenges. *International Journal of Enhanced Research in Science Technology & Engineering, 11*, 65–79.

Al-Emran, M., & Griffy-Brown, C. (2023). The role of technology adoption in sustainable development: Overview, opportunities, challenges, and future research agendas. *Technology in Society, 73*, 102240.

Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics* (pp. 13–22). ACL.

Allouhi, A., Rehman, S., Buker, M. S., & Said, Z. (2023). Recent technical approaches for improving energy efficiency and sustainability of PV and PV-T systems: A comprehensive review. *Sustainable Energy Technologies and Assessments, 56*, 103026.

Ampornphan, P., & Tongngam, S. (2020). Exploring technology influencers from patent data using association rule mining and social network analysis. *Information, 11*(6), 333–352.

Andreoni, A., Chang, H.-J., & Labrunie, M. (2021). Natura non facit saltus: Challenges and opportunities for digital industrialisation across developing countries. *The European Journal of Development Research, 33*, 330–370.

Cammarano, A., Varriale, V., Michelino, F., & Caputo, M. (2023). Employing online big data and patent statistics to examine the relationship between end product's perceived quality and components' technological features. *Technology in Society, 73*, 102231.

Catelli, R., Bevilacqua, L., Mariniello, N., di Carlo, V. S., Magaldi, M., Fujita, H., & Esposito, M. (2022). Cross lingual transfer learning for sentiment analysis of Italian TripAdvisor reviews. *Expert Systems with Applications, 209*, 118246.

Chand Bhatt, P., Kumar, V., Lu, T.-C., & Daim, T. (2021). Technology convergence assessment: Case of blockchain within the IR 4.0 platform. *Technology in Society, 67*, 101709.

Cho, J., & Lee, J. (2013). Development of a new technology product evaluation model for assessing commercialization opportunities using Delphi method and fuzzy AHP approach. *Expert Systems with Applications, 40*(13), 5314–5330.

Cho, R.L.-T., Liu, J. S., & Ho, M.H.-C. (2021). The development of autonomous driving technology: Perspectives from patent citation analysis. *Transport Reviews, 41*(5), 685–711.

da Silva Barboza, A., Aitken-Saavedra, J. P., Ferreira, M. L., Aranha, A. M. F., & Lund, R. G. (2021). Are propolis extracts potential pharmacological agents in human oral health?-A scoping review and technology prospecting. *Journal of Ethnopharmacology, 271*, 113846.

Dai, T., Xiao, Y., Liang, X., Li, Q., & Li, T. (2022). ICS-SVM: A user retweet prediction method for hot topics based on improved SVM. *Digital Communications and Networks, 8*(2), 186–193.

de Paulo, A. F., de Oliveira Graeff, C. F., & Porto, G. S. (2023). Uncovering emerging photovoltaic technologies based on patent analysis. *World Patent Information, 73*, 102181.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171–4186.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *Proceedings of the 32nd International Conference on International Conference on Machine Learning, 37*, 1180–1189.

Gui, M. Z., & Xu, X. G. (2021). Technology forecasting using deep learning neural network: Taking the case of robotics. *IEEE Access, 9*, 53306–53316.

Han, J. Y., Chen, Y. C., & Li, S. Y. (2022). Utilising high-fidelity 3D building model for analysing the rooftop solar photovoltaic potential in urban areas. *Solar Energy, 235*, 187–199.

Jaffe, A. B., & de Rassenfosse, G. (2017). Patent citation data in social science research: Overview and best practices. *Journal of the Association for Information Science and Technology, 68*(6), 1360–1374.

Jiang, X. R., & Liu, J. J. (2023). Extracting the evolutionary backbone of scientific domains: The semantic main path network analysis approach based on citation context analysis. *Journal of the Association for Information Science and Technology, 74*(5), 546–569.

Jin, Q. C., Jiang, J., Li, J. C., & Yang, K. W. (2020). Emerging technology identification and selection based on data-driven: Taking the unmanned systems as an example. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 1180–1189). IEEE.

Kang, I., Yang, J., Lee, W., Seo, E. Y., & Lee, D. H. (2023). Delineating development trends of nanotechnology in the semiconductor industry: Focusing on the relationship between science and technology by employing structural topic model. *Technology in Society, 74*, 102326.

Khan, J., Ullah, I., & Yuan, J. Y. (2022). $CsPbI_3$ perovskite quantum dot solar cells: Opportunities, progress and challenges. *Materials Advances, 3*(4), 1931–1952.

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. *BMC Medical Imaging, 22*(1), 69.

Klevorick, A. K., Levin, R. C., Nelson, R. R., & Winter, S. G. (1995). On the sources and significance of interindustry differences in technological opportunities. *Research Policy, 24*(2), 185–205.

Kumar, R. V. K., Ukko, J., Rantala, T., & Saunila, M. (2024). The value of novel technologies in context to performance measurement and management: A systematic review and future research directions. *Data and Information Management, 8*(1), 100054. https://doi.org/10.1016/j.dim.2023.100054

Kwon, K., Jun, S., Lee, Y.-J., Choi, S., & Lee, C. (2022). Logistics technology forecasting framework using patent analysis for technology roadmap. *Sustainability, 14*(9), 54–84.

Lee, C. (2021). A review of data analytics in technological forecasting. *Technological Forecasting and Social Change, 166*, Article 120646.

Lee, J., & Sohn, S. Y. (2021). Recommendation system for technology convergence opportunities based on self-supervised representation learning. *Scientometrics, 126*(1), 1–25.

Lee, Y., Kim, S. Y., Song, I., Park, Y., & Shin, J. (2014). Technology opportunity identification customized to the technological capability of SMEs through two-stage patent analysis. *Scientometrics, 100*(1), 227–244.

Li, X., Zhang, J., & Ouyang, J. (2019). Dirichlet multinomial mixture with variational manifold regularization: Topic modeling over short texts. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 7884–7891.

Ma, Y., Kong, L., Lin, C., & Yang, X. (2022). Research on the identification of generic technology of eco-friendly materials based on text mining. *Environmental Science and Pollution Research, 29*(23), 35269–35283.

Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Doshi, A., & Yang, Y. (2022). RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology Artificial Intelligence, 4*(5), e210315.

Miao, Y., Yu, L., & Blunsom, P. (2015). Neural variational inference for text processing. *ArXiv, abs/1511.06038*.

Ministry of Science and Technology of China (MOST). (2012). Solar Power Development "Twelfth Five-Year" Plan. Retrieved from https://www.gov.cn/zwgk/2012-04/24/content_2121638.htm

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *Profesional De La Información, 29*(1), 4–24.

National Science & Technology Council of America (NSTC). (2018). Strategy for american leadership in advanced manufacturing. Retrieved from https://www.manufacturingusa.com/reports/strategy-american-leadership-advanced-manufacturing

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. (pp. 100–108). Association for Computational Linguistics.

Noh, H., Song, Y.-K., & Lee, S. (2016). Identifying emerging core technologies for the future: Case study of patents published by leading telecommunication organizations. *Telecommunications Policy, 40*(10–11), 956–970.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9–33.

Ren, H. Y., & Zhao, Y. H. (2021). Technology opportunity discovery based on constructing, evaluating, and searching knowledge networks. *Technovation, 101*, 102196.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). Association for Computing Machinery

Royer, J. M. (1979). Theories of the transfer of learning. *Educational Psychologist, 14*, 53–69.

Schaller, A.-A., Vatananan-Thesenvitz, R., Pulsiri, N., & Schaller, A.-M. (2019). The rise of digital business models: An analysis of the knowledge base. In *2019 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 1–13). IEEE.

Schmid, M. (2023). Revisiting the definition of solar cell generations. *Advanced Optical Materials, 11*(20), 2300697.

Shen, Y.-C., Wang, M.-Y., & Yang, Y.-C. (2020). Discovering the potential opportunities of scientific advancement and technological innovation: A case study of smart health monitoring technology. *Technological Forecasting and Social Change, 160*, Article 120225.

Son, C., Kim, J., & Kim, Y. (2020). Developing scenario-based technology roadmap in the big data era: An utilisation of fuzzy cognitive map and text mining techniques. *Technology Analysis & Strategic Management, 32*(3), 272–291.

Song, B., & Suh, Y. (2019). Identifying convergence fields and technologies for industrial safety: LDA-based network analysis. *Technological Forecasting and Social Change, 138*, 115–126.

Talebpour, M., De Herrera, A. G. S., & Jameel, S. (2023). Topics in contextualised attention embeddings. In the 45th European Conference on Information Retrieval (ECIR) (pp. 221–238). Springer-Verlag

U.S. Department of energy (DOE). (2016). SunShot 2030. Retrieved from https://www.energy.gov/eere/solar/sunshot-2030

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.

Wan, X. J., Chen, F., Li, H. L., & Lin, W. B. (2022). Potentially related commodity discovery based on link prediction. *Mathematics, 10*(19), 3713.

Wang, J., Ding, Z., Liu, Z., & Feng, L. (2024). Technology opportunity discovery based on patent analysis: A hybrid approach of subject-action-object and generative topographic mapping. *Technology Analysis & Strategic Management, 36*(9), 2070–2083.

Wang, N., Gong, Z., Xu, Z., Liu, Z., & Han, Y. (2021). A quantitative investigation of the technological innovation in large construction companies. *Technology in Society, 65*, 101533.

Wang, R., Zhou, D., & He, Y. (2019). ATM: Adversarial-neural topic model. *Information Processing & Management, 56*(6), 102098.

Yang, J. L., Chen, X. F., Huang, C. Y., & Ma, T. M. (2022). Hotspot and frontier discovery of hydrogen detection technology based on bibliometrics. *Sensor Review, 42*(5), 599–610.

Yoon, B., & Magee, C. L. (2018). Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. *Technological Forecasting and Social Change, 132*, 105–117.

Yu, S., Han, R., Zheng, Y., & Gong, C. (2022a). An Integrated AMPSO-CLSTM Model for Photovoltaic Power Generation Prediction. *Frontiers in Energy Research, 10*, 815256. https://doi.org/10.3389/fenrg.2022.815256

Yu, X., Wang, J., Hong, Q. Q., Teku, R., Wang, S. H., & Zhang, Y. D. (2022b). Transfer learning for medical images analyses: A survey. *Neurocomputing, 489*, 230–254.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43–76.