Scientometrics (2023) 128:5883–5904
https://doi.org/10.1007/s11192-023-04819-x

Identification of emerging technology topics (ETTs) using BERT-based model and sematic analysis: a perspective of multiple-field characteristics of patented inventions (MFCOPIs)

Bowen Song[1] · **Chunjuan Luan**[1,2] · **Danni Liang**[1]

Received: 15 March 2022 / Accepted: 16 August 2023 / Published online: 3 September 2023

Abstract
The proliferation of large language models (LLMs) has significantly expanded the landscape of research on technology opportunity identification. However, there remains a crucial need to enhance the accuracy and interpretability of results obtained through emerging technology topic identification. In this paper, we present a novel approach that leverages a BERT-based model and semantic analysis to identify emerging technology topics (ETTs) from the perspective of multiple-field characteristics of patented inventions (MFCOPIs). By utilizing a unique dataset encompassing MFCOPI, our methodology emphasizes an increased proportion of novel technical processes in the analysis content while mitigating the interference of redundant technical information. To enhance the interpretability of recognition results, our proposed model employs the BERT model for detecting potential content similarities in inventive characteristics and incorporates semantic structure analysis to expand the technical process content. We empirically validate our model by employing nanotechnology as a case study, demonstrating its effectiveness and accuracy. Through our research, we extend the existing methodologies for recognizing emerging technology, ultimately elevating the quality of recognition results.

Keywords Emerging technology topics · Inventive characteristics · Multiple fields · BERT · Sematic analysis · Nanotechnology

Introduction

In this paper, we aim at proposing a novel approach to identify emerging technology topics more accurately and efficiently, by employing a combination method of deep learning and sematic analysis. Emerging technologies play a pivotal role in shaping the modern economic landscape, acting as a crucial catalyst for transformation and innovation across

✉ Bowen Song
bowensong333@163.com

[1] Institute of Humanities & Social Sciences, Dalian University of Technology, Dalian 116085, China

[2] School of Intellectual Property, Dalian University of Technology, Panjin 124221, China

diverse sectors. By facilitating the development of pioneering models, they bolster the innovative capacity of enterprises, thus engendering a lively and dynamic economic environment. (Abernathy & Utterback, 1978). There is no unified view of emerging technology in different fields, but the characteristics of "High risk" (Rotolo et al., 2015), "High influence" (Marsili, 2001), and "Creative destruction" (Christensen & Raynor, 2013) have been widely recognized. The discernment of emerging technology topics is paramount to uncovering the vanguard trends in science and technology, ensuring equitable distribution of national resources, and maximizing the output of scientific research institutions via collaboration with technological enterprises.

The primary objective of identifying emerging technology topics is to detect and scrutinize novel terms, assess potential technology trajectories, and forecast the innovation focal points of future technologies. Nonetheless, certain limitations persist in the research methodologies. (1) Prevailing research typically regards keywords as the identification results of emerging technology topics; however, a solitary technical term or an aggregation of technical keywords and ancillary information can readily obfuscate expert analysis and interpretation, consequently impacting the accuracy of identification outcomes. For example, metformin and aspirin drugs exhibit disparate technical effects as technology evolves. The presence of a singular named entity without semantic connections may engender ambiguity in recognition results. (2) Technical term recognition research often neglects emerging phenomena, such as cross-domain technology applications, substantial enhancements in technical performance, or the adaptive development of mature technology. Consequently, it fails to pinpoint emerging technology topics that materialize due to shifts in these characteristics.

Drawing on the similarity of technical features, this study introduces a recognition model that integrates deep learning and semantic analysis, striving to augment the precision and efficiency of emerging technology topic identification outcomes and mitigate the challenges associated with interpreting such results. The framework of the identification model is structured as follows: First, we extract the patent title, technical novelty, application, and advantage information from patent data, subsequently constructing the inventive characteristics set; Second, employing the BERT pre-training model in deep learning, we train the existing inventive characteristics and vectorize recent inventive characteristics; Third, we extract keyword performance using a combination of semantic and aggregation techniques, ensuring consistency in keyword performance; Fourth, we refine the themes of emerging technologies; Lastly, we employ topic coherence to assess the recognition efficacy of the model. Utilizing nanotechnology as a case study, we conduct emerging technology topic recognition to validate the effectiveness of the model's recognition capabilities.

Literature review and theoretical background

Definition of "emerging technologies topics"

Emerging technology is a scientific innovation, which may establish a new industry or transform an old industry (Day et al., 2000). Emerging technology topics are commonly employed as a comprehensive descriptor for a class of nascent technological endeavors. Despite the resemblance in nomenclature, emerging technology and emerging technology topics are frequently conflated in research, yet they possess notable distinctions in terms of their concepts and features. Adopting the perspective proposed by Reardon, these

differences can be categorized into three facets. (Furukawat, 2015; Reardon & Sara, 2014; Yoon et al., 2011): (1) An emerging technology topic embodies the consolidation of attributes unique to innovative technologies, signifying a novel classification of technology; (2) Conceptually, the topic of technology is clear, with no ambiguity and uncertainty in emerging technology; (3) Divergent from emerging technology, the amalgamation of emerging technology topics requires a temporal accumulation process, during which the formation of topics steers the advancement of subsequent technologies. Integrating prevailing perspectives, this study delineates emerging technology topics as keywords or phrases capable of encapsulating the shared novelty attributes of emerging technologies over time.

Text-based emerging technology topics identification

Academics have extensively investigated methods for effectively and accurately detecting, mining, and identifying emerging technology topics from patents, documents, and other sources of information. The expert evaluation approach based on expert experience and knowledge is one of the methods commonly used by scholars and mainly refers to the use of the Delphi method (Yun et al., 1991), expert consultation, Scenario Planning (Tseng. et al., 2009), and other methods. Utilizing a scoring system, these methods identify the technical subject with the highest composite score as the emerging technology topic, predicated on expert subjective comprehension and evaluative analysis. The TRIZ method is also used in an attempt to discover emerging technologies by tracing the source of technical problems (Al′tshuller, 1999). While the aforementioned qualitative analysis has yielded favorable outcomes for long-term forecasting, the comprehension of technological topic novelty is frequently constrained by the disciplinary background and technical expertise of the experts involved.

The index analysis method is another method used frequently in finding characteristics of technology. It refers to the construction of a comprehensive index system based on rapid growth (Kleinberg, 2003), novelty (Porter & Detampel, 1995), convergence (Song & Luan, 2019), coherence (Rotolo et al., 2015), wide impact (Wang, 2018), comprehensive characteristics (Zhang et al., 2017), and other characteristics of emerging technology to identify the theme of emerging technology (Zhang et al., 2021). Frequently employed indicators encompass patent application quantity, patent word frequency increment, patent citation increment, convergence index, economic risk index, and technology opportunity analysis. The technical disparities between fields may influence the index evaluation process, consequently leading to substantial fluctuations in the identification outcomes. Furthermore, trend extrapolation analysis (such as S-curve, Hype Cycle, and technology life cycle) and other indicator-based methodologies are extensively employed in emerging technology topic identification research.

Index analysis is not the sole quantitative approach for topic identification in emerging technologies. Relationship network analysis, grounded in citation networks (Citation, Co-citation, Bibliographic Coupling), co-occurrence networks, and text data similarity networks, are also frequently employed in the identification of emerging technology topics. As shown in Fig. 1, relationship network analysis generally includes five steps, Step 1: Text collection; Step 2: Node Determination; Step 3: Building Relationships; Step 4: Network analysis and Step 5: Topic identification. For example, the cosine similarity relation of the patent reference coupling network is used to mine the emerging technologies in the outlier state (Song et al., 2018); patent citation networks are used to identify new technology clusters (Érdi et al., 2013), and keyword co-occurrence network and dynamic recursive neural

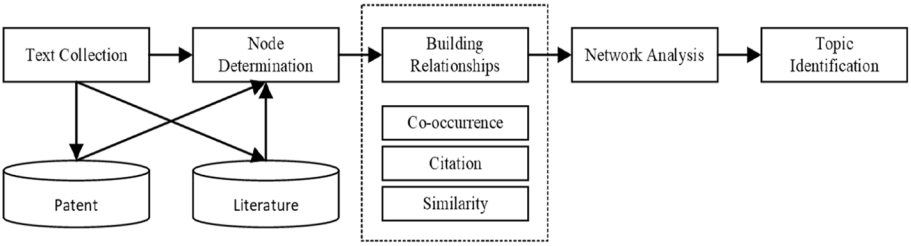

Fig. 1 General process of relationship network analysis

network (DRNN) are combined to predict potential technology trends (Choudhury et al., 2020). Relational network analysis is not used solely for the identification of technical topics, cluster analysis is often used as a tool for refining technical topics. Though the network structure-based computation method offers increased accuracy and can precisely depict the spatiotemporal evolution of patents or technical topics through similarity relationships, it encounters issues such as limited interpretability and an inability to convey technical disparities in depth.

Machine learning processing in emerging technology identification

As natural language processing technology advances, topic recognition models incorporating machine learning have garnered considerable interest in recent years. By integrating machine learning and deep learning techniques, these models have substantially enhanced the depth of text content relationship analysis at the semantic level. Consequently, they have increased the accuracy of semantic similarity calculations and more precisely captured the relationships between complex indicators and outcomes.

The majority of enhanced recognition models continue to adhere to the research concepts of relationship network analysis, incorporating machine learning at various stages of the analytical process to boost result accuracy and recognition efficiency. To mitigate the impact of limited data on recognition outcomes, Yuan Zhou proposes a predictive approach that combines data augmentation and deep learning (Zhou et al., 2020). Building upon network nodes and relationships, Florian examines the case of service robots, integrating machine learning and the Support Vector Machine (SVM) model for the analysis of emerging field information. (Kreuchauff & Korzinov, 2017). Hassan etc. put forward a new topic recognition model combining citation analysis and deep learning by using the sample data of 64 dimensional indicators (Hassan et al., 2018). Blei and Lafferty apply a Latent Dirichlet Allocation (LDA) model to monitor the dynamics of temporal changes of emerging trends (Blei & Lafferty, 2006). Based on Subject-Action-Object (SAO) semantic vector analysis, Park et al. argue an outlier patent recognition method using patent semantic similarity (Park et al., 2012). Professor A conducted a comprehensive comparison of seven analytical models—LSTM, NNAR, LightGBM, linear regression, polynomial regression, Escore, and Naive Method—under both global and local strategies to assess their predictive accuracy. In the end, LSTM was selected, combined with nine quantitative metrics, to forecast the changing heat of technology theme development (Liang et al., 2021). The analysis aimed to identify innovative emerging technology topics with promising potential. Doc2vec, Word2vec, BERT, GPT and other machine learning methods have also been applied to the recognition model to improve the granularity of the recognition results.

PageRank, Centrality, Word Frequency and other indicators are still commonly used in network analysis (Joung & Kim, 2017), but the object of measurement is changing from node to vector.

Incorporating machine learning does enhance the recognition models' accuracy; however, a simple low similarity relationship between topics may not sufficiently represent an emerging technology topic. The selection of novel content for emerging technology profoundly influences the outcomes. Simultaneously, the interpretability of the results remains low. For instance, without the subject words' semantic background information, discerning the specific meaning that the clustering aims to convey solely through keyword information proves challenging. Consequently, further research is required to address the limitations in the semantic expression of external features within related models.

Theoretical discussions on relationship of inventive characteristics and emerging technology topics

The identification of emerging technology topics emphasizes the "precision" and "comprehensiveness" of the results. Technology topics should not only reflect the uniqueness of inventive characteristics, exploring novel aspects of things, but also distill inventive characteristics that share common or similar properties based on the inherent characteristics of the field. Despite the richness of current research on technology themes, there is a lack of exploration into the sources of inventive characteristics. Hence, this section begins by reviewing the pertinent literature on inventive characteristics and delves into the relationship between inventive characteristics and emerging technology topics from three angles: definition, measurement, and connection.

The definition of "inventive characteristics"

The stipulation of inventive characteristics serves the purpose of mitigating duplication in quantifying the output of inventive endeavors, as it is evident that two inventions identical in their characteristics do not constitute two distinct contributions to the reservoir of technological knowledge. Since 1960s, inventive characteristics have emerged as an active role in the research on innovation management (Kuznets, 1962). Scholars have advanced diverse viewpoints concerning the fundamental nature of innovative characteristics. These perspectives encompass new processes (Teece, 1986), novel technological combinations (Brockhoff, 1992), application of knowledge across domains (Nelson 1985), processes exceeding current thresholds (Podolny & Stuart, 1995), major and radical shifts occur in the knowledge structures (Hayashi, 2004), efficiency-enhancing technologies, and energy-conserving technologies (Kuznets, 1962). In recent years, new perspectives have been proposed. Özel and Pénin proposed that inventive characteristics, from the perspective of patent licensing, represent the exclusivity of a technology (Özel & Pénin, 2016). Jaffe et al. combined the citation patterns of patents and proposed that inventive characteristics represent unique and high-quality attributes that extend beyond the cumulative nature of related research, benefiting business development (Jaffe & De Rassenfosse, 2017). These alterations, arising from diverse aspects such as knowledge foundations, technological amalgamations, materials, processes, and efficiency, collectively fall under the purview of "inventive characteristics," irrespective of whether they entail incremental enhancements or radical transformations.

The measurement of "inventive characteristics"

In 1957, Schmookler analyzed the correlation between the number of technologists and the quantity of patents (Schmookler, 1957). He discovered the significant role of potential inventors in promoting technological innovation and provided detailed explanations of the specific professions (chemists, architects, designers, draftsmen, assayers, et al.) associated with technologists. Inspired by Schmookler, early scholars attempted to measure inventive characteristics by exploring data variations related to inventors, the quantity of patents (Ernst, 2003), publications (Furukawat, 2015), trademark (Mendonça et al., 2004), and other sources. With the rise of text analysis, the exploration of inventive characteristics is no longer limited to changes in the quantity of innovation carriers. Scholars have begun to ponder how to conduct in-depth mining from the content of the text. ST&I data (Porter & Detampel, 1995), keywords (Kleinberg, 2003; Chen, 2004), topics (Zhang et al, 2016), abstracts (Ma et al., 2021), and full patent texts (Choi et al., 2011) are all considered crucial sources of inventive characteristics.

The connection of "inventive characteristics" and "emerging technology topics"

The shared trait between "inventive characteristics" and "emerging technology topics" lies in their capacity to exemplify unique and high-caliber attributes of technological innovation. Furthermore, disparities exist between the two. "Inventive characteristics" offer an elaborate account of distinctive attributes within technology innovation, whereas "emerging technology topics" place greater emphasis on concisely summarizing similar innovative attributes of the same kind (Gerken & Moehrle, 2012; Lee, 2021). Looking at it from the perspective of the topic identification process, we need to extract "innovative characteristics" from a vast amount of patent texts to better represent potential novel information and thereby enhance the efficiency of identification. The extraction of "emerging technology topics" necessitates the categorization of innovative characteristics sharing similar attributes.

This study posits that the significance of innovative characteristics in identifying emerging technology topics lies primarily in the following aspects: Most importantly, innovative characteristics are the refinement of new processes, new materials, new combinations, etc. in patent texts. Stripping features from common attributes can improve the reflection of emerging technology topics on new content. Moreover, an emerging technology topic embodies the consolidation of attributes unique to innovative technologies, signifying a novel classification of technology (Reardon & Sara, 2014). Leveraging the similarity between innovative points, we can accomplish the categorization of innovative characteristics and, thereby, enhance the extraction of their unique properties more effectively. Lastly, innovative characteristic is an improvement in machine learning performance. The ultimate goal of innovative characteristics is to obtain superior data, enabling learning algorithms to extract patterns and achieve better outcomes.

In light of the aforementioned challenges, this study, grounded in a systematic analysis of the composition principles of emerging technologies, integrates deep learning and SAO semantic analysis to put forth a novel approach for emerging topic recognition.

The primary contributions of this work can be summarized as follows:

(1) By integrating the two prevalent innovative modes of emerging technologies, disruptive and sustaining innovation, a selection strategy for innovative characteristic is presented. This approach aims to refine the data source of emerging technology topics and enhance the accuracy of identification outcomes at the data-collection stage.
(2) The BERT model is integrated to reinforce the training of prevailing semantic relationships, comprehensively learning the terminology, word associations, and sentence structure within technical texts. This method heightens the recognition model's ability to discover hidden semantic connections.
(3) Semantic analysis technology is applied to tag process information within technology texts, extracting the subject-action-object structure corresponding to the technology topic. This approach enhances the interpretability of the topic recognition results.

Methodology

Research framework

In an effort to elucidate the novelty feature content and feature source of emerging technology topics, this study presents an identification model for emerging technology topics predicated on the similarity of technical features. The model is specifically designed to address the innovation pathways and innovative characteristics within these emerging topics. The overall process is divided into five steps: (1) data collection and innovative characteristics extraction; (2) model training based on innovative characteristics; (3) vectorization of novel innovative characteristics; (4) the construction of multi-dimensional innovative characteristics similarity network, and the k-means algorithm to cluster the characteristics; (5) extract the core keywords and key semantic structures in the clustering features. The detailed process is shown in Fig. 2.

Data collecting and innovative characteristics processing

Patent data is an important data source for identifying emerging technology topics (Song et al., 2018). The choice of the patent database will directly affect the accuracy of the final

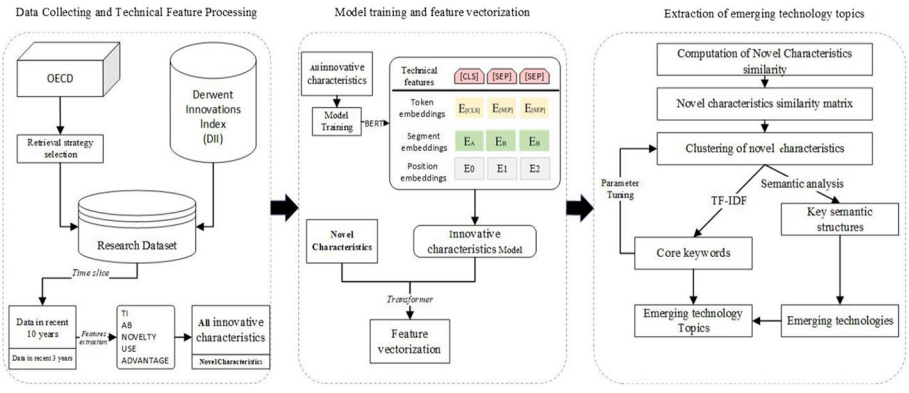

Fig. 2 Overview of emerging technology topic recognition model design

identification results. Derwent Innovations Index (Abbr. as DII) includes the data of more than 50 patent-granting agencies around the world, covering 96% of patent data to ensure that obtaining technical features is comprehensive. Each Derwent patent data contains information such as title, inventor, patent number, classification number, and summary among others, which accurately locate the theme information.

Time slice selection

The choice of technology cycle significantly influences both model identification and technological novelty. Early scholars generally believed that a technology takes approximately 15 years from its initial stage to its recession (Christensen & Raynor, 2013). However, due to increased competition, emerging technologies currently face a more intense market environment than traditional technologies. From a perspective of technology benefits, the development cycle of emerging technologies has been shortened to 5–10 years, of which 3–5 years are to form early concepts that are key to the theme of emerging technologies (Porter et al., 2002). Gartner considers that in the current environment, characterized by abundant media coverage, commercial hype, and public discussions, a new concept requires 2–5 years to transition from its nascent stage to a phase of rapid expansion.

Considering that a 5-year timeframe is a critical milestone for the formation of emerging concepts, we have chosen to extract emerging technology topics from patent texts within the past 5 years. Simultaneously, we selected patent data from the 10-year period preceding this as background information for identifying emerging technology topics, making it easier for machine learning models to understand common terminology, combinations, and expressions within the domain. It is important to note that the recent 5-year patent data used for predictions is not included in the 10-year training data set.

Innovative characteristics extraction

Innovative characteristic extraction aims to distill relevant attributes of emerging technology topics from patent text, mitigating the influence of superfluous features and unrelated characteristics on recognition outcomes. Pertaining to the specific information encompassed by emerging technologies, various perspectives have been put forth by scholars (Adner & Levinthal, 2002; Christensen & Raynor, 2013). Rotolo et al. portrays emerging technologies as radical innovation, possessing the ability to swiftly aggregate copious resources and display relatively expeditious growth characteristics within a condensed time frame (Rotolo et al., 2015). Day et al. proposes a more encompassing definition, positing that emerging technologies not only comprise disruptive technological innovations originating from radical transformations but also encompass continuous innovative technologies formed through the convergence of multiple, previously independent research outcomes (Day et al., 2000). Drawing from the aforementioned perspectives, two scenarios emerge as focal points for innovative characteristic extraction of emerging technology topics: disruptive innovation and incremental innovation. Disruptive innovation frequently triggers considerable shifts in technology themes, primarily manifesting as alterations in technical terminology, such as novel technological terms or the adoption of distinct performance attributes. Conversely, incremental innovation maintains the prevailing technological term environment but incites modifications in technical combinations, including application domains and implementation strategies.

In light of the variations in required technical feature extraction content under distinct innovation modes, a more comprehensive innovative characteristic extraction strategy has been selected. To capture term alterations resulting from disruptive innovation, corresponding changes are identified by examining patent title fields and novelty fields within the patent text. A complete patent text contains a large number of fields: patent number, patent title, inventor, patent ownership agency, abstract, patent publication date, etc. Among them, the patent title (TI) field is the most concise overview of the content and application of the patented invention (Derwent, 2023). The novelty (NOVRLTY) field emphasizes the technical features and contents of this patent that are different from other patents. on the other hand, to capture technical combination changes stemming from incremental innovation, corresponding alterations are identified by examining patent title fields, use fields, and advantage fields within the patent text. The technical use (USE) field describes the application method, field, technical scheme, and other information of the technology in detail, while the technical advantage (ADVANTAGE) field reflects the change and improvement in the new technology performance to the existing technology. These fields containing innovative characteristic information provide us with a Multiple-field characteristics of patented inventions (MFCOPI). This helps us to more accurately grasp the changes in emerging technologies from various dimensions, such as materials, processes, technological combinations, and knowledge foundations. Table 1 summarizes the technical field tags and meanings of each part (WIPO, 2023). The above description items can reflect emerging information overall, including emerging technology topics under different innovation paths, and are the key data representing the innovative characteristics.

Upon confirming the innovative characteristics in the patent text, the Python natural language processing tool is used to extract the innovative characteristic information of each patent in the training data set and the recent data set, respectively, including the technical terms, performance characteristics, and application fields. Other information in the innovative characteristic is restored and sorted by word form to obtain the complete and standardized innovative characteristics. Depending on the data source, the characteristics extracted from the past 5 years are referred to as “novel characteristics”, while those extracted from a decade-long patent collection are termed “background characteristics”.

BERT model training and feature vectorization

Existing technology is the core reference for evaluating the theme of emerging technology (Lee et al., 2018; Zhang et al., 2018). Compared to the traditional word frequency analysis and semantic recognition analysis, the introduction of the deep learning model can not only improve the recognition of new technical terms, but also further analyze the semantic

Table 1 Innovative characteristic fields and meanings

Features fields	Field tags	Meaning
Topic	TI	Reveal the invention content and novelty of the patent
Abstract	AB	A detailed description of the technical scheme, novelty, and application field
Technical novelty	Novelty	New feature information that does not belong to the existing technology
Technical use	Use	Application mode or scheme of technical features
Technical advantage	Advantage	Change or improvement of new technology performance

meaning of words, word order and the relationship between sentences in technical features. In addition, it could carry out the vectorization coding of recent innovative characteristics and more accurately locate the changes of application fields, technical performance, technology portfolio, and other information under different development paths.

Deep learning pre-trained model

In this study, BERT's (Bidirectional Encoder Representations from Transformers) pre-trained model is used for the training of background characteristics in the training dataset. This is a language-processing deep learning encoder proposed by Jacob Devlin (Devlin et al., 2018; Vaswani et al., 2017). The BERT model adopts the dual mechanism of Mask Language Model (MLM) and Next Sentence Prediction (NSP) in the training of technical features. That is, through the MLM strategy, this model expands the field of vision by randomly covering some words, and achieving two-way training of technical characteristics (Taylor, 1953). The NSP strategy is used to contrast learning to get the sentence-level semantic representation and to understand the correlation between sentences.

Therefore, the background characteristics of each input BERT training model will be divided into three parts. This will then be analyzed at the three levels for words, semantics, and syntax. The first layer is token embedding, which contains the semantic information of technical terms, keywords, and other words in technical features. The second layer is segment embedding, which contains the sequence information between different statements in the technical features. The third layer is position embedding, which includes the word-order relationship between words in innovative characteristics. The deep learning of existing innovative characteristics can be achieved through the BERT pre-trained model, which can characterize the existing technical attributes. As shown in Fig. 3, the innovative characteristic, "Sensor composites gas sensing layer that board placed on substrate," indicates a new sensor operation process. In the token embedding layer, the technical features are generated into word-embedding vectors according to the word content; in the segment embedding layer, the innovative characteristics are embedded into sentences according to the special identifier [SEP]; in the position embedding layer, the word-order information is embedded according to the word order. The application of the BERT model improves the effect of vectorization of innovative

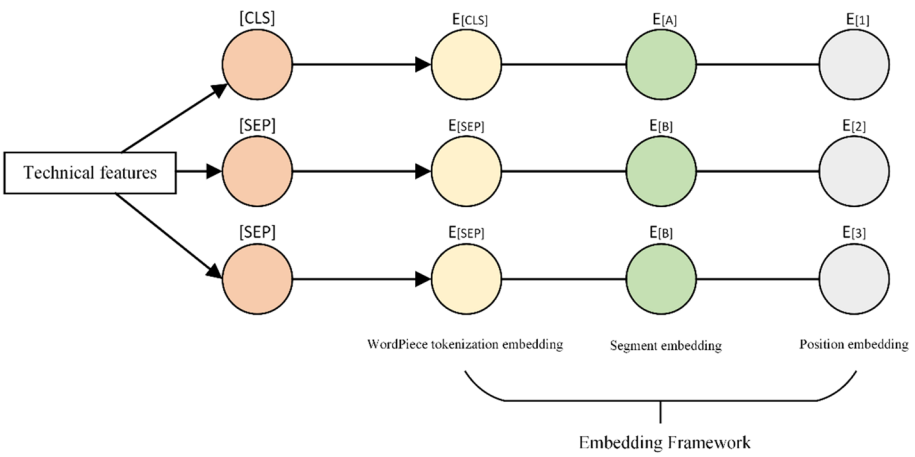

Fig. 3 BERT pre-trained model input architecture

characteristics (Devlin et al., 2018). As a coder based on deep learning, compared with the traditional method based on word frequency, the model not only considers the meaning of the word itself, but also considers the correlation between the word and the sentence, and realizes the full description of the existing innovative characteristics.

Vectorization of novel characteristics

Because the importance of words is different, we can not only rely on word frequency to judge the word's importance, so the determination of word weight in the model is also an important issue in feature analysis. The BERT model introduces the multi-head self-attention mechanism (Vaswani et al., 2017; Chen et al., 2021); Through repeated measurement of word-sentence relevance and vector dimension, it can accurately reflect the weight of technical keywords, as shown in Eq. (1).

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) * V \tag{1}$$

Q, K, V represent the composition matrix of the query (Q) vector, keys (K) vector, and values (V) vector of the technical keyword respectively; QK^T indicates the degree of relevance between keywords and sentences; d_k represents the number of Q and K vector dimensions. After calculation, we can obtain the vector representation of each weighted technical keyword.

$$ET_{emb} = \text{feed}\left(W_z * MultiHead(Q, K, V) + X_{emb}\right) + X_{emb} \tag{2}$$

As shown in Eq. (2), ET_{emb} is the final coding result of the BERT pre-training model for novel characteristics. Feed(x) represents the forward propagation process, W_z represents the weight matrix of technical features, $MultiHead(Q, K, V)$ indicates the result of the long attention mechanism, and X_{emb} represents the recent technical feature input matrix. Through the above process, the n-dimensional vector of each technical keyword mapping is finally obtained.

Emerging technology topic identification

Emerging technology topics are not only the extraction of technical terms nor terms and other named entities in innovative characteristics but also the overall induction of similar technical problems and their respective solutions. Based on the technical feature vector, this study uses the "Clustering + Keyword + Semantics" pattern to enhance the accuracy of technical subject recognition results. Among them, the K-means clustering algorithm condenses innovative characteristics with the same technical structure, preparation process, and performance after vectorization. Keyword analysis can extract the core technical information, SAO semantic analysis can extract the key concepts and structures in the innovative characteristics.

Novel characteristics clustering

K-means algorithm is an unsupervised clustering algorithm, which iteratively and repeatedly calculates the clustering center until the result converges (Macqueen, 1967). It can also classify the system according to the distance between nodes. The cosine distance is usually selected to measure the distance between nodes in clustering. The calculation process is

shown in Eq. (3). It indicates the cosine similarity between the novel characteristics of cosα; the higher the cosine value, the higher the similarity between the two vectors.

$$\cos\alpha = \frac{\sum_{i=1}^{n}\left(A_i * B_i\right)}{\sqrt{\sum_{i=1}^{n}\left(A_i\right)^2} * \sqrt{\sum_{i=1}^{n}\left(B_i\right)^2}} = \frac{x_A}{x_A} * \frac{x_B}{x_B} \tag{3}$$

$$S_i = \frac{b_i - a_i}{\max\left(b_i, a_i\right)} \tag{4}$$

In this study, the Silhouette coefficient is used as the judgment basis for the selection of clustering results. Silhouette coefficient (S_i) is the comprehensive measure of the degree of separation and cohesion in the clustering results, as shown in the calculation formula in Eq. (4). S_i represents the Silhouette coefficient, which ranges from—1 to 1. The closer the coefficient is to 1, the more obvious the clustering effect. a_i represents the average distance between feature I and the same type of technical feature, and b_i represents the average distance between feature I and other types of innovative characteristics. The selection of clustering numbers will be judged based on the optimal solution of Silhouette coefficient.

Semantic analysis of novel characteristics

The extraction of technical topics usually focuses on the mining of technical terms in clustering; however, it is difficult for scholars to rely on keyword information to accurately judge emerging technology trends. Therefore, this study introduces semantic analysis to help scholars understand the meaning of topics. In Subject-Action-Object (SAO), the S stands for solutions while the A-O stands for technical problems. SAO semantic structure can more fully reflect the application information and conceptual structure of technology compared to the limited information that words can express (Park et al., 2012; Yoon et al., 2011, 2021). The combination of triple structure can represent the actions between subject and object, and other functional information in technical features. The technology function identification analysis of the SAO semantic structure has been widely used in the research of technology evolution analysis, technology opportunity analysis, knowledge organization identification, and technology road mapping.

Therefore, in this study, we first employ the Stanford Parser natural language processing tool to label the parts of speech in the clusters. Subsequently, we utilize natural language processing tools to extract semantic structures containing Subject-Action-Object from the labeled innovative characteristics within the clusters. Finally, we consider the frequently occurring semantic structures in each cluster as helpful tools for interpreting the technology topics. By extracting the core structure information of technology combination, which is product efficacy and performance contained in the extraction results in the semantic structure, we can more clearly control the technical content conveyed by the subject clustering.

Keywords extraction of novel characteristics

Under technical topics, common extraction tools include the term frequency-inverse document frequency (TF-IDF) model, TextRank model, LDA topic model, and other fusion models. To better understand the content combination between technical topics and the

above-mentioned emerging technologies, TF-IDF combined with expert opinions is used to realize topic recognition. The TF-IDF algorithm measures the weight of each word in the technical features to highlight the importance of representative terms in the technical features. The word weight increases proportionately to the frequency of occurrence in the technical features and decreases in inverse proportion to the frequency of occurrence in the word bag model. The calculation formula is as shown in Eq. (5).

$$TF * IDF = \frac{n_{i,j}}{\sum_k n_{k,j}} * log \frac{|D|}{\left|j : t_i \in d_j\right|} \tag{5}$$

Cluster analysis forms a large number of technical features with the same attributes into a feature set. Each feature set contains a large number of keywords that reflect the attributes of the set, measure the keyword weight through the TF-IDF algorithm, and extract the keywords with representative innovative characteristics.

Empirical analysis and result

Nanotechnology refers to research on the composition of substances at the scale of 1 to 100 nm. This mainly includes the creation of new substances by manipulating and rearranging atoms and molecules (Zhang et al., 2013). The rapid development of nanotechnology has improved people's understanding of material composition and performance. It has also made human micro perception and control ability reach an unprecedented level. Therefore, in the proposed report by the United States, "Converging Technologies for Improving Human Performance" (Roco & Bainbridge, 2002), nanotechnology is regarded as a fundamental element of future technological innovation.

Data collection of nanotechnology field

This study takes the Derwent innovation index (DII) as the data source and retrieves and collects the patent data related to nanotechnology from 2008 to 2022 as the analysis data set. It then selects the latest nanotechnology retrieval strategy released by the Organization for Economic Co-operation and Development (OECD), constructs the nanotechnology retrieval strategy combined with the characteristics of the International Patent Classification (IPC). The search strategy employed in this investigation is IP = (B82B OR B82Y OR B82B-001/00 OR B82B-003/00 OR B82Y-005/00 OR B82Y-010/00 OR B82Y-015/00 OR B82Y-020/00 OR B82Y-025/00 OR B82Y-030/00 OR B82Y-035/00 OR B82Y-040/00 OR B82Y-099/00). A total of 45,276 patent texts filed between 2018 and 2022 were retrieved and compiled to serve as the prediction dataset for topic identification, whereas 55,493 patent texts filed between 2008 and 2017 were utilized as the training dataset.

Nanotechnology pre-trained model

In accordance with the construction process of the emerging technology topic recognition model, we extracted innovative characteristics from 55,493 patents filed between 2008 and 2017, and subsequently built a training dataset for nanotechnology. Next, we extracted innovative characteristics from 45,276 patents filed between 2018 and 2022, creating a prediction dataset. The Pre-trained model is selected to enable deep learning of the feature

set of training technology. The three levels of word semantic, sentence relation, and word order relation are then extracted to analyze the mask language model. The BERT model includes two coding layers: the mask language model obtained from the training set, which is used to encode the new technical feature set. 12 heads are selected by the multi-head, self-attention mechanism, and each new technical feature is mapped into a 768-dimensional vector to obtain the vectorized new technical feature.

Novel characteristics clustering

The pairwise distance between 44,671 recent technical features is calculated by cosine similarity, and a network containing K random centroids is built with the technical features as the nodes. Use the cosine similarity value as the node distance, we conduct K-means clustering analysis, set the value of clustering coefficient K between 2 and 10, and observe the change of the Silhouette coefficient by adjusting the value of K. Given the susceptibility of the K-means algorithm to initial value selection, variations in initial node assignments for identical K values can impact both clustering outcomes and iteration counts. Consequently, this investigation employed a stringent repetitive experimental approach to pinpoint the optimal K value. In particular, 30 independent clustering analyses were conducted for each unique K value, with subsequent computation of the associated silhouette coefficients. The analysis of these silhouette coefficient distributions facilitated the systematic determination of the suitable number of topic categories for the dataset in question.

Figure 4 displays the distribution of silhouette coefficients for clustering the innovative characteristics of nanotechnology, we employed a box plot to illustrate the distribution relationship between K values and silhouette coefficients. Examining the overall trend of silhouette coefficient variations, within the range of K values from 2 to 10, the coefficients demonstrate an initial increase followed by a decline. When K is set to 6, the maximum, minimum, median, and quartile values of the silhouette coefficients surpass those of other results. Concurrently, in terms of box length, the data exhibits reduced volatility and relatively stable clustering outcomes when K is assigned a value of 6. Hence, we can preliminarily determine that when the number of clusters is set to 6, the distinctiveness between innovative characteristics reaches its maximum value.

We reduced the innovative characteristics vector space to three dimensions when K is set to 6 for visualization purposes, as depicted in Fig. 5. Dots with different colors represent the latest innovative characteristics of different contents, and the distance between dots indicates the similarity between technical features. The clustering results are represented by groups of corresponding color dots. The clearer the boundary between different color clusters, the more obvious the clustering effect. Therefore, the clustering analysis based on BERT has a very obvious effect on the clustering of recent technical features.

K-means cluster analysis divides nanotechnology into six categories according to the degree of similarity of the technical features. They are named C1, C2, C3, C4, C5 and C6, respectively. C1 contains 16,480 innovative characteristics accounting for 36.4% of the total novel characteristics; C2 contains 5750 innovative characteristics, which accounts for 12.7% of the total novel characteristics; C3 contains 6881 innovative characteristics, accounting for 15.2% of the total novel characteristics; C4 contains 5478 innovative characteristics, accounting for 12.1% of the total novel characteristics; C5 contains 5931 innovative characteristics, accounting for 13.1% of the total novel characteristics; C6 contains 4753 innovative characteristics, accounting for 10.5% of the total novel characteristics. As a result, C1 contains the highest number of innovative characteristics, significantly

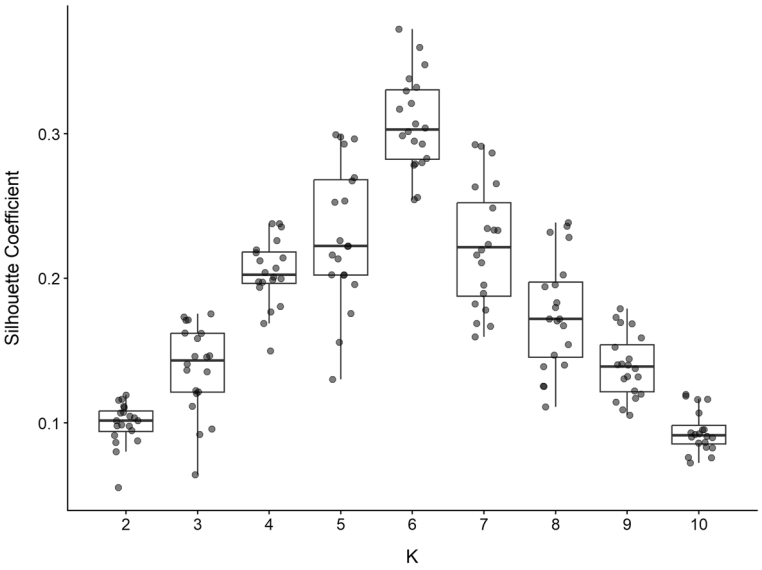

Fig. 4 Silhouette coefficient distribution

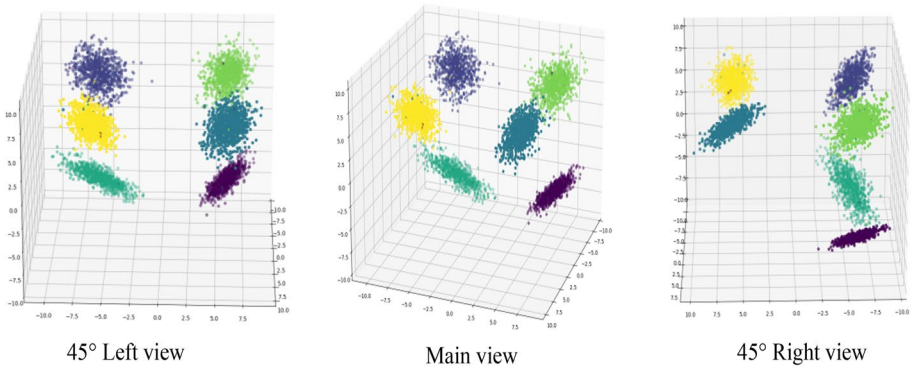

Fig. 5 Clustering results of novel characteristics

surpassing those in other clusters, while C6 encompasses the fewest innovative characteristics. The number of innovative characteristics in the remaining clusters shows minimal variation.

Semantic analysis of novel characteristics

Based on the above steps, the novel characteristics in the six clusters are semantically marked by using the Stanford parser natural language processing tool. After completing the semantic labeling of the six clustering novel characteristics, we extract the major subject action-object semantic structure, and high-frequency keywords using the TF-IDF algorithm, as shown in Table 2.

The high-frequency keywords involved in Cluster 1 include electromagnetic; beam; luminescent; emission and spectrum. These keywords are commonly used terms in nanodevices. After narrowing down the scope of topics, we combined the information from the semantic structure "The method involves scanning probe lithography" to comprehend the key keyword information. By analyzing the terms "scanning probe lithography" and "involve" within the context, we identified a class of emerging technology related to scanning probe lithography that emerged in 2020. Similarly, with the assistance of the semantic structure "The integrated drive has directional couplers" and relevant patents, we discovered another category of technology related to integrating directional couplers into driver chips or modules to facilitate wireless communication. Furthermore, by analyzing the semantic structure "A nanowire structure connects the source/drain structure", we have identified that the emerging technology topics also involves optimization techniques related to nanowire-based transistors. Using the same approach of combining keywords with semantic analysis, we analyzed the remaining five clusters. The high-frequency keywords involved in Cluster 2 include separate; purify; raw; sieve; buffer; iron; impurity and residue. The emerging technology applications mentioned in the semantic structures include "novel preparation process for nanoscale composite materials", "preparation of nanoscale ferric chloride materials", and "nanomaterial functionalization techniques". These technologies are all nano-fabrication technologies emerging in 2019. The high-frequency keywords involved in Cluster 3 include target; polypeptide; tumor; carrier; cell; lipid; drug and plasma. The emerging technology applications mentioned in the semantic structures include "extraordinary molecular motor-enabling technology for DNA-packaged motors", "extraction technology of medicinal ingredients", and "preparation technology of bifunctional hybrid thin film". From the semantic structure, a series of nanotechnology applications in the biopharmaceutical domain can be observed. The high-frequency keywords involved in Cluster 4 include particle; doped; nm; epitaxial; sputtering; deposition and nanowires. The emerging technology applications mentioned in the semantic structures include "mass spectrometry techniques for characterizing complex mixtures and determining molecular structures", "High-performance liquid chromatography (HPLC)", and "technologies to enhance thermoelectric conversion efficiency". The high-frequency keywords involved in Cluster 5 include catalyst; palladium; platinum; gold; sulfur; molar and perovskite. The emerging technology applications mentioned in the semantic structures include "multi-quantum dot chip technology", "improving mechanical properties, enhancing absorption capacity, and increasing surface area through the use of nanopowders' porous structure", and "composite catalyst preparation technology". The high-frequency keywords involved in Cluster 6 include sinter; calcine; heat; furnace; energy; battery and cathode. The emerging technology applications mentioned in the semantic structures include "catalytic converters to reduce pollutant emissions" and "reduce negative properties of solid propellants".

Emerging technology topic identification

In combination with the high-frequency keywords appearing in Cluster 1, we can position this cluster as "Nano-devices and Integrated Technology". By matching the high-frequency SAO semantic structures with patents, we identified that the emerging technologies under topic "Nanodevices and Integrated Technology" primarily involve the reformation and innovation of technologies in areas "Nanoparticle sensor technology" and "Nano-photonics

Table 2 Overall information of nanotechnology cluster analysis

Cluster	Topics	S (subject)	A (action)	O (object)
C1	Electromagnetic, beam, luminescent, emission, spectrum	The method	Iinvolve	Scanning probe lithography
		The integrated drive	Have	Directional couplers
		A nanowire structure	Connect	The source/drain structure
C2	Separate, purify, raw, sieve, buffer, iron, impurity, residue	Extract	Synthesize	Nanomaterial
		Nano-sheet	Comprise	Iron chloride
		A nanomaterial	Modify	The fluorinated carbon electrode material
C3	Target, polypeptide, tumor, carrier, cell, lipid, drug, plasma	Bacteriophage DNA packaging motor	Modify	Portal protein
		Anticancer composition	Comprise	Simarouba glauca extract
		Dual-functional hybrid film	Containing	Ferrocene-doped cobalt-based metal
C4	Particle, doped, nm, epitaxial, sputtering, deposition, nanowires	The method	Characterize	Mass spectrometry
		The test solution	Inject	High performance liquid chromatograph
		nano-thermoelectric	Use	Nano-thermoelectric material micro-zone
C5	Catalyst, palladium, platinum, gold, sulfur, molar, perovskite	Gold island	Enhance	Multiple quantum dot chip
		Nano-powder	Contain	Porous structure
		Titanium dioxide	Composite	Composite catalyst
C6	Sinter, calcine, heat, furnace, energy, battery, cathode	Catalytic converter	Remove	Pollutant
		Energy material	Decompose	Pollutant
		Solid propellant	Prepare	The nano-high-energy material

and photonics in nanotechnology". Finally, we can acquire the emerging technology topics in the field of nanotechnology between the period 2018–2022, as shown in Fig. 6.

Other emerging technology topics and the prominent emerging technologies they encompass are described as follows: The topic of Cluster 2 is "Nanomaterials Preparation and Synthesis", which involves emerging technologies such as "Nanomaterial separation and purification" and "Nanomaterial synthesis techniques". The topic of Cluster 3 is "Nanobiotechnology", which involves emerging technologies such as "Biosensing technology" and "Targeted drug delivery technology". The topic for Cluster 4 is "Nano-characterization and Detection", with "Atomic force microscopy" and "Nano-mechanical testing techniques" serving as representative emerging technologies under this theme. The topic for Cluster 5 is "Nano-catalysts and Surface Science", and the emerging technologies within this theme include "Design of nano-catalysts" and "Cross-disciplinary Applications of Nano-catalysts". The title for Cluster 6 is "Nano-energy and Environmental Technologies", featuring emerging technologies primarily pertaining to aspects "Nanobattery Technology" and "Nanoscale heat treatment technology".

Model checking

Parameter comparison

Considering the subjectivity of expert interpretation, this study introduces topic coherence to test the effect of topic clustering and compares the topic recognition model based on BERT with LDA model and Word2vec model. Coherence measures whether words in the same topic are coherent. In this paper, the UCI measure method proposed by Newman is used to segment the subject words based on the sliding window, and the topic coherence of the model is obtained by calculating the point state mutual information of all word pairs in a given word (Newman et al., 2011). The value range of topic coherence is [0,1]. The closer the value is to 1, the more obvious the effect.

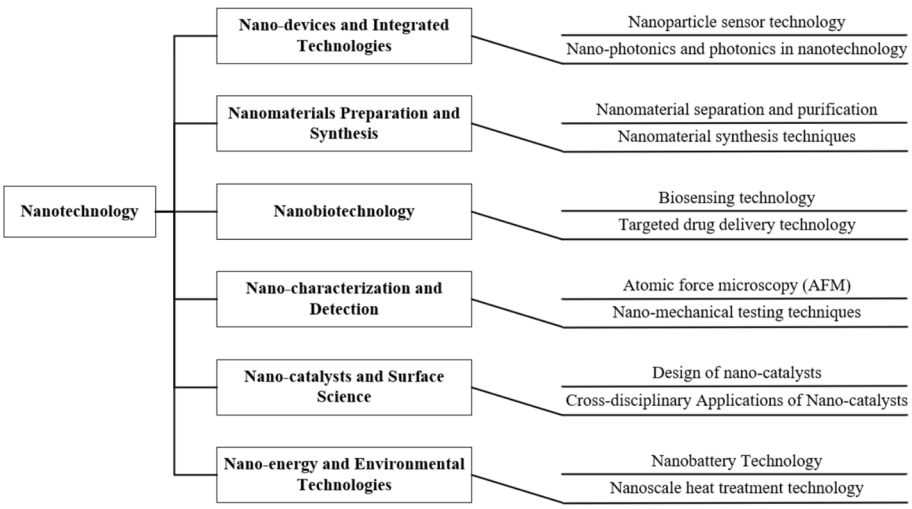

Fig. 6 Emerging technology topics in Nanotechnology

As shown in Table 3, The coherence of the LDA model is 0.411, and the coherence of the Word2vec model is 0.437. The coherence of the BERT based recognition model is 0.503, which is higher than the other models. It shows that compared with the general topic recognition model, the emerging technology topic recognition model integrating Word2vec, and K-means clustering can be more accurate and concise, and the distinction between technology topics is more obvious. At the same time, since the model introduces semantic analysis on the basis of topic keyword recognition, it improves the interpretability of emerging technology secondary topics and further reduces analysis error.

Identification effect verification

Compare the identification results with the key R & D plans of nanotechnology in the "14th Five-Year Plan (2021–2025)" released by China in 2021. The topics of "Nano-devices and Integrated Technology", "Nanomaterials Preparation and Synthesis", "Nanobiotechnology", "Nano-characterization and Detection", "Nano-catalysts and Surface Science" and "Nano-energy and Environmental Technologies" have appeared in the scientific and technological frontier research projects. The emerging technologies further identified through semantic analysis are also found among the 28 prioritized categories of cutting-edge nanotechnology research supported by the leading organizations. It further verifies that the identification results of emerging technology topics in this study are consistent with the actual development.

Discussion and conclusions

The formation and change of emerging themes in technology have a dramatic impact on scientific and technological progress, economic development, social change, and many others. The early identification and monitoring of emerging technology topics are very important for the rational allocation of national resources. It is also necessary for optimal output of scientific research institutions and scientific and technological enterprises. Considering this, the study aims at the lack of sensitivity and interpretability of various technological innovation paths in the current emerging technology topic recognition. This study also proposes an emerging technology topic recognition model based on the similarity of innovative characteristics. First, the model extracts the patent title, novelty, technical use, and technical advantage information from the patent data. It then constructs the background characteristics set and the novel characteristics set according to the time distribution. The deep learning algorithm is used to train the background characteristics set, fully identifying the words, semantics, word order, sentences, and other information in the existing innovative characteristics. Moreover, it vectorizes the novel characteristics, the fusion of feature vector and clustering algorithm, thus realizing the clustering of innovative characteristics with the same attribute. Using technology keyword extraction to mine core topics, combined with semantic analysis to extract the core semantic structure of innovative characteristics, we can highlight the content of technology combination, efficacy, and performance. Taking nanotechnology as an example, the emerging technology topic recognition results show that the model can clearly and accurately mine the

Table 3 Topic coherence comparison of different models

Evaluation coefficient	BERT + K-means	LDA	Word2vec
Coherence	0.503	0.411	0.437

emerging technology topic information in the target field and describe the emerging technology content involved in the topic, at the same time. Through the comparison with other agent recognition models, it is confirmed that the effect is better than LDA, Word2vec and other machine learning models.

By combining deep learning and semantic analysis, this study identifies an emerging technologies prediction method, which is an important supplement to the research perspective of existing methods. In the selection of innovative characteristics, this study makes full use of the text information of patents and strengthens the hidden content in the technology that is not easily detected. In the process of topic condensation, this study introduces a deep learning algorithm to achieve a more effective similarity calculation method. In the prediction of emerging technology topics, the combination of words and semantics is used. By obtaining technical keywords, the emerging technology topics are more clearly explained by associating relevant semantic information. This research method has more advantages in recognition efficiency, wide application fields and interpretability of recognition results.

Still, the method in this paper has some limitations. First, we only select patent data as the basis for subject identification of emerging technologies, mainly for the technology driven prediction of emerging technologies. In the future, we will consider introducing scientific and technological literature data to further improve the accuracy of identification results theoretically. Second, this research model only selects the Derwent database as the data source of new features. In the future, the research scope will be expanded, and corresponding feature extraction strategies will be designed according to the features of data from different data sources.

Funding Funding was provided by National Natural Science Foundation of China (Grant Nos. 71774020, 71473028).

References

Abernathy, W. J., & Utterback, J. M. (1978). Patterns of industrial innovation. *Technology Review, 80*(7), 40–47.

Adner, R., & Levinthal, D. A. (2002). The emergence of emerging technologies. *California Management Review, 45*(1), 50–66.

Al′tshuller, G. S. (1999). *The innovation algorithm: TRIZ, systematic innovation and technical creativity*. Technical Innovation Center Inc.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Paper presented at the machine learning, proceedings of the twenty-third international conference* (*ICML 2006*), Pittsburgh, Pennsylvania, USA, June 25–29, 2006.

Brockhoff, K. (1992). Instruments for patent data analyses in business firms. *Technovation, 12*(1), 41–59.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences, 101*(1), 5303–5310.

Chen, J., Jiang, S., Wang, M., Xie, X., & Su, X. (2021). Self-assembled dual-emissive nanoprobe with metal-organic frameworks as scaffolds for enhanced ascorbic acid and ascorbate oxidase sensing. *Sensors and Actuators B: Chemical, 339*, 129910.

Choi, S., Yoon, J., Kim, K., Lee, J. Y., & Kim, C. H. (2011). SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. *Scientometrics, 88*(3), 863–883.

Choudhury, N., Faisal, F., & Khushi, M. (2020). Mining temporal evolution of knowledge graphs and genealogical features for literature-based discovery prediction. *Journal of Informetrics, 14*(3), 101057.

Christensen, C., & Raynor, M. (2013). *The innovator's solution: Creating and sustaining successful growth*. Harvard Business Review Press.

Day, G., Schoemaker, P., & Gunther, R. E. (2000). *Wharton on managing emerging technologies*. Wiley.

Derwent. (2023). Derwent innovations index: Derwent innovations index user guide. Retrieved from https://clarivate.com/webofsciencegroup/support/wos/dii/.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., & Zalányi, L. (2013). Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics, 95*(1), 225–242.

Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information, 25*(3), 233–242.

Furukawat, M. (2015). Identifying the evolutionary process of emerging technologies: A chronological network analysis of World Wide Web conference sessions. *Technological Forecasting and Social Change, 2015*(91), 280–294.

Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. *Scientometrics, 91*(3), 645–670.

Hassan, S. U., Imran, M., Iqbal, S., Aljohani, N. R., & Nawaz, R. (2018). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics, 117*(3), 1645–1662.

Hayashi, A. M. (2004). Technology trajectories and the birth of new industries: Markets develop according to the specific paths by which innovations in a given field occur. *MIT Sloan Management Review, 45*(3), 7–9.

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.

Joung, J., & Kim, K. (2017). Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. *Technological Forecasting and Social Change, 114*, 281–292.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery, 7*(4), 373–397.

Kreuchauff, F., & Korzinov, V. (2017). A patent search strategy based on machine learning for the emerging field of service robotics. *Scientometrics, 111*(2), 743–772.

Kuznets, S. (1962). Inventive activity: Problems of definition and measurement. *The rate and direction of inventive activity: Economic and social factors* (pp. 19–52). Princeton University Press.

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

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change, 127*, 291–303.

Liang, Z., Mao, J., Lu, K., Ba, Z., & Li, G. (2021). Combining deep neural network and bibliometric indicator for emerging research topic prediction. *Information Processing & Management, 58*(5), 102611.

Ma, T., Zhou, X., Liu, J., Lou, Z., Hua, Z., & Wang, R. (2021). Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies. *Technological Forecasting and Social Change, 173*, 121159.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 14*(1), 281–297.

Marsili, O. (2001). *The anatomy and evolution of industries: Technological change and industrial dynamics*. Edward Elgar Publishing.

Mendonça, S., Pereira, T. S., & Godinho, M. M. (2004). Trademarks as an indicator of innovation and industrial change. *Research Policy, 33*(9), 1385–1404.

Nelson, R. R. (1985). *An evolutionary theory of economic change*. Harvard University Press.

Newman, D., Bonilla, E. V., & Buntine, W. (2011). Improving topic coherence with regularized topic models. *Advances in Neural Information Processing Systems*, 24.

Özel, S. Ö., & Pénin, J. (2016). Exclusive or open? An economic analysis of university intellectual property patenting and licensing strategies. *Journal of Innovation Economics Management, 21*(3), 133–153.

Park, H., Yoon, J., & Kim, K. (2012). Identifying patent infringement using SAO based semantic technological similarities. *Scientometrics, 90*(2), 515–529.

Podolny, J. M., & Stuart, T. E. (1995). A role-based ecology of technological change. *American Journal of Sociology, 100*(5), 1224–1260.

Porter, A. L., & Detampel, M. J. (1995). Technology opportunities analysis. *Technological Forecasting and Social Change, 49*(3), 237–255.

Porter, A. L., Roessner, J. D., Jin, X. Y., & Newman, N. C. (2002). Measuring national 'emerging technology' capabilities. *Science and Public Policy, 29*(3), 189–200.

Reardon, S. (2014). Text-mining offers clues to success. *Nature, 509*(7501), 410.

Roco, M. C., & Bainbridge, W. S. (2002). Converging technologies for improving human performance: Integrating from the nanoscale. *Journal of Nanoparticle Research, 4*(4), 281–295.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology. *Research Policy, 44*(10), 1827–1843.
Schmookler, J. (1957). Inventors past and present. *The Review of Economics and Statistics, 39*(3), 321–333.
Song, B. W., & Luan, C. J. (2019). Impact indicator on measuring multi-dimension technological convergence. In *17th international conference on scientometrics & informetrics* (*ISSI2019*) (Vol. I).
Song, K., Kim, K., & Lee, S. (2018). Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents. *Technological Forecasting and Social Change, 128*, 118–132.
Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*(4), 415–433.
Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy, 15*(6), 285–305.
Tseng, F. M., Cheng, A. C., & Peng, Y. N. (2009). Assessing market penetration combining scenario analysis, Delphi, and the technological substitution model: The case of the OLED TV market. *Technological Forecasting and Social Change, 76*(7), 897–909.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the Association for Information Science and Technology, 69*(2), 290–304.
WIPO. (2023). Guidelines for the wording of titles of inventions in the patent documents. Retrieved from https://www.wipo.int/export/sites/www/standards/en/pdf/03-15-01.pdf.
Yoon, B., Kim, S., Kim, S., & Seol, H. (2021). Doc2vec-based link prediction approach using SAO structures: Application to patent network. *Scientometrics*, 1–30.
Yoon, J., Choi, S., & Kim, K. (2011). Invention property-function network analysis of patents: A case of silicon-based thin film solar cells. *Scientometrics, 86*(3), 687–703.
Yun, Y., Jeonger, G. H., & Kim, S. H. (1991). A Delphi technology forecasting approach using a semi-Markov concept. *Technological Forecasting and Social Change, 40*(3), 273–287.
Zhang, R., Zhang, Y., Dong, Z. C., Jiang, S., Zhang, C., Chen, L. G., Zhang, L., Liao, Y., Aizpurua, J., Luo, Y. E., & Yang, J. L. (2013). Chemical mapping of a single molecule by plasmon-enhanced Raman scattering. *Nature, 498*(7452), 82–86.
Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099–1117.
Zhang, Y., Wu, M., Miao, W., Huang, L., & Lu, J. (2021). Bi-layer network analytics: A methodology for characterizing emerging general-purpose technologies. *Journal of Informetrics, 15*(4), 101202.
Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting and Social Change, 105*, 179–191.
Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology, 68*(8), 1925–1939.
Zhou, Y., Dong, F., Liu, Y., Li, Z., Du, J., & Zhang, L. (2020). Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics, 123*(1), 1–29.