Automated Identification of Emerging Technologies: Open Data Approach

Ljiljana Dolamic[1,†], Julian Jang-Jaccard[1,*,†], Alain Mermoud[1,†] and Vincent Lenders[1,†]

[1]Cyber-Defence Campus, armasuisse Science and Technology, Thun, Switzerland

Abstract

Identifying emerging technologies and forecasting their trends is pivotal for stakeholders and decision-makers across academia, industry, and government agencies. The current strategies employed to track technology trends often rely on proprietary closed datasets and often rely on the insights of human domain experts. Not only are these approaches expensive and manual, but they are also time-consuming. In this study, we introduce an automated method for identifying emerging trends through a quantitative approach that utilizes extensive publicly available data, including patents, publications, and Wikipedia Pageview statistics. Our method proposes four criteria – novelty, growth, impact, and coherence – to automatically score technologies, based on a mathematical foundation. This approach enables the monitoring of tech trends across various sectors in an automated manner, without the need for domain experts. The results obtained through rigorous evaluation, benchmarked against similar reports from leading market research firms, illustrate a low recall rate paired with high precision, affirming the reliability of our proposed method. Furthermore, our method identifies emerging technologies not present in similar market reports, highlighting its unique capabilities.

Keywords

technology monitoring, emerging technologies, attributes of emergence, scientometrics, open source data, machine learning, informetrics, natural language processing

1. Introduction

Understanding emerging technologies is crucial for various entities, including industry, academia, and government agencies. It can shape strategic decisions, improve competitive positions, and create opportunities for technology strategies. Owing to these considerations, there is a substantial need for identifying emerging technologies, prompting widespread media coverage on the topic and leading market research firms like Gartner and Forrester to offer services promising deeper insights.

Despite the common and widespread use of the term 'emerging technologies,' there is no single standard agreement on what constitutes the term. This lack of a clear definition makes it challenging to develop a scientifically sound methodology to identify emerging technologies. Gartner's renowned Hype Cycle for Emerging Technologies, while intuitive, cannot serve as an underlying model and has faced criticism in the literature for being considered unscientific, inconsistent, generic, and subjective [1]. Other market research firms, such as Forrester and IHS Markit, also produce annual reports on emerging technologies, yet the methodology for identifying these technologies remains unclear.

Research in the area of identifying emerging technologies primarily relies on qualitative methods, expert systems, and survey-based approaches. For quantitative methods, researchers have utilized open datasets and S-curve models to identify technology emergence [2, 3, 4, 5]. S-Curve models, based on logistic or Gompertz growth concepts, provide a solid mathematical foundation. However, most studies focus on specific predetermined sets of technologies, making it challenging to devise a general method for identifying emerging technologies [6].

In this paper, we introduce a novel approach for identifying emerging technologies based on their coverage in publicly available data sources, including patents, publications, and Wikipedia Pageview statistics. Unlike previous studies, we have not preselected any specific set of technologies. Our method is transparent, does not require expert input, and gives reproducible results for any technology.

The remainder of this paper is organized as follows: Section 2 provides a survey of existing research. In Section 3, we offer a description of the data used. Section 4 outlines the proposed methodology. We present the evaluation results in Section 5. The limitation of our proposed method is discussed in Section 6. Finally, Section 7 concludes the paper with future work.

2. Related Work

Definitions for the term 'emerging technologies' in the literature often overlap but are based on distinct characteristics. For example, some authors (e.g., [7, 8, 9, 10, 11]) emphasize the potential impact of the technology on the economy or society, covering both evolutionary change and disruptive innovations. Others, like Boon [12], prioritize uncertainty about a technology's future evolution. Some researchers combine both potential and uncertainty aspects [13, 14], while others underline novelty and growth [15].

The myriad of characteristics chosen to define emerging technologies has given rise to diverse scientometric approaches for measurement [16, 17], lacking a standardized definition of the underlying concept of emergence. A comprehensive analysis by Rotolo, Hicks, and Martin [18] explores existing research on the definition of emerging technologies, aggregating comparable approaches. They identify five main characteristics—radical novelty, relatively fast growth, coherence, prominent impact, and uncertainty—commonly appearing across the studied research. We adopt this definition as a foundational framework for

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), April 23 24, 2024, Changchun, China and Online

*Corresponding author.

†These authors contributed equally.

ljiljana.dolamic@ar.admin.ch (L. Dolamic); julian.jang-jaccard@ar.admin.ch (J. Jang-Jaccard); alain.mermoud@ar.admin.ch (A. Mermoud); vincent.lenders@ar.admin.ch (V. Lenders)

0000-0002-0656-5315 (L. Dolamic); 0000-0002-1002-057X (J. Jang-Jaccard); 0000-0001-6471-772X (A. Mermoud); 0000-0002-2289-3722 (V. Lenders)

our study.

Predicting emerging technologies often relies on publicly available datasets, commonly leveraging patents such as those from the United States Patent and Trademark Office (USPTO), Global Patent Index (GPI), and Thompson Innovation. Numerous publications advocate for the use of bibliometric methods to extract data and identify emerging technologies, followed by deploying growth models for prediction. In the work of Daim et al. [19], bibliometric methods, US patent analysis, and S-curves were employed for forecasting technologies such as fuel cells, food safety, and optical storage. Similarly, Ranaei et al. [3] used expert interviews to fit data acquired by text-mining patents into growth curve models for predicting hybrid cars and fuel cells. Text-mining on patents and fitting to S-curves were also proposed in [20], and Bengisu et al. [21] found correlations between patent and publication data extracted by scientometric methods for 20 technologies, deploying S-curves for forecasting. S-Curve models for predicting emerging technologies were also proposed by [2, 22].

In recent times, artificial intelligence has regained significant attention, leading to the use of machine learning to model and predict emerging technologies. Kyebambe and Hwang [23, 24] employed supervised learning on citation graphs from USPTO data to automatically label and forecast emerging technologies. Similarly, Zhou [25] applied supervised deep learning on worldwide patent data, with training sets labeled based on Gartner's Hype Cycle.

3. Data

We primarily use three different datasets: patent data from USPTO, publication data from arXiv, and statistical data from Wikipedia Pageviews.

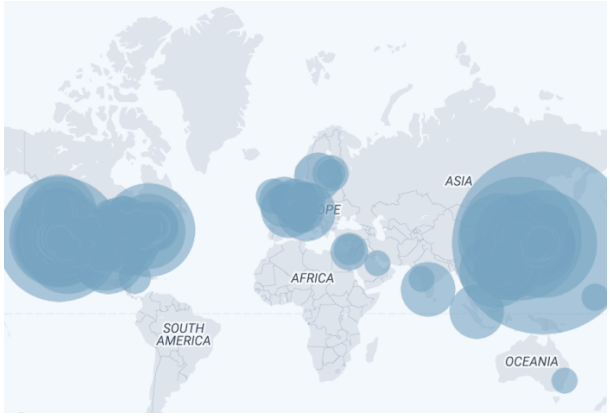

Figure 1: Top 200 locations by patent count for granted patents during 2013 - 2023 (Source from [26])

Patents from PatentsView[1]**:** Patent information provides valuable insights into the latest innovations, trends, and competitive landscapes within various industries. We utilize PatentsView to acquire patent information from the USPTO for granted patents since 1976. As of December 5, 2023, there are over 8 million records of granted patents available for free download for further analysis. Figure 1 provides a glimpse of the top 200 locations worldwide for patents granted by the USPTO since 2013. We utilize a subset of around 6.6 million patent records for our study.

Publications from arXiv[2]**:** We employ arXiv as a primary publication source, taking advantage on its free distribution model for open-access scholarly articles. The repository hosts over 2.4 million publications spanning computer science and diverse scientific disciplines since 1993. Figure 2 displays the number of submissions to arXiv since August 1991. Our study focuses on a subset of approximately 1.4 million arXiv publications.

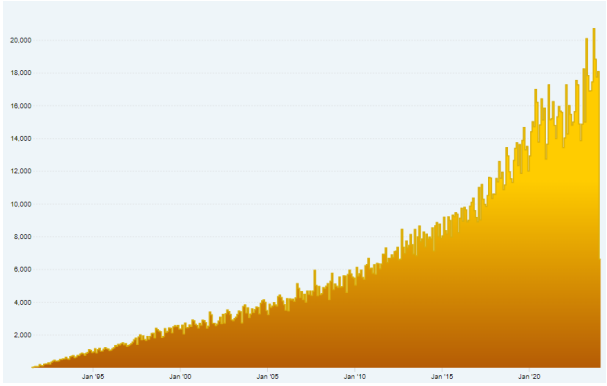

Figure 2: Number of arXiv submissions since 1991 (Source from [27])

Wikipedia Pageview Statistics [3]**:** In addition, we incorporate Wikipedia Pageview statistics which indicates the number of visitors to a Wikipedia article within a specified time frame. This offers insight into real-time public interest and engagement, serving as a dynamic and accessible indicator of emerging trends and technologies. Figure 3 illustrates an example of a monthly pageview statistics for the keyword 'deep learning'.

Leveraging the Wikipedia API, we retrieved the monthly views for 50,954 articles relevant to the technology.

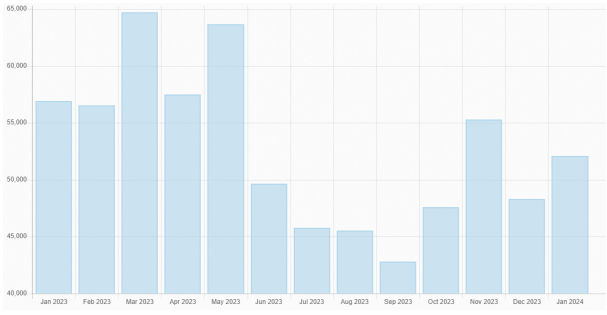

Figure 3: Number of Pageviews of the topic 'deep learning' during Jan 2023 - Jan 2024 (Source from [28])

4. Methodology

In this section, we outline our methodology, and Figure 4 offers a comprehensive overview of the entire process.

The proposed method is initiated by classifying each Wikipedia article as either technology-related or not, em-

[1]https://patentsview.org/

[2]https://arxiv.org/

[3]https://en.wikipedia.org/wiki/Wikipedia:Pageview_statistics

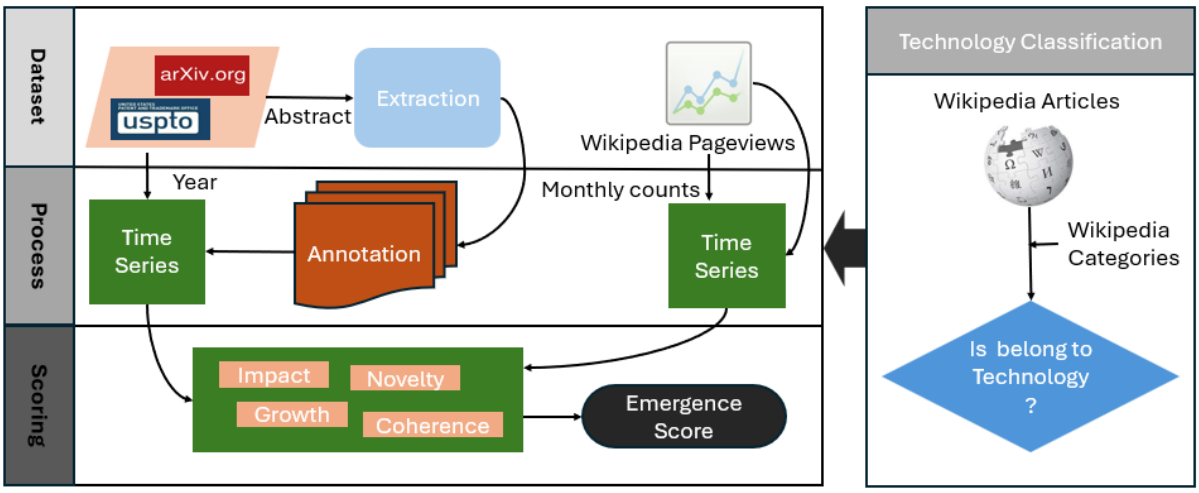

Figure 4: Overview of the Proposed Methodology

ploying a binary classification approach termed as technology classification.

Once this classification is established, we extract abstracts from USPTO and scholarly arXiv publications. These abstracts undergo annotation using the DBPedia tool [4], aligning the text with Wikipedia articles. This annotation process aims to link the abstract content to relevant Wikipedia entries. To reduce noise, we eliminate annotations occurring fewer than 5 times and those not aligned with the technology classification.

The resulting filtered annotations, all within the technology classification, serve as the basis for constructing time series. The count of mentions for each technology $t \in T$ per year is summed across each data source $d \in D$, reflecting the increasing occurrences of patents and publications over time. Mathematically, this can be represented as:

$$\text{Total Count}(t) = \sum_{d \in D} \text{count}(t, d)$$

where $\text{count}(t, d)$ is the count of mentions for technology t in data source d. We then compute relative counts in relation to the total number of technology mentions per year, represented as:

$$\text{Relative Count}(t) = \frac{\text{Total Count}(t)}{\text{Total Technology Mentions per Year}}$$

Furthermore, monthly Wikipedia Pageviews are obtained for all technologies and transformed into time series. These time series, along with Wikipedia categories, contribute to the computation of four scores—Novelty, Growth, Impact, and Coherence—each derived from the definitions provided by [18]. Finally, we aggregate and normalize these four scores to generate an emergence score for each technology.

4.1. Technology Classification

The output of annotated abstracts from patents and publications contains noise, as each annotation refers to a Wikipedia article, not necessarily related to technology. To address this issue, we devised a two-step methodology named 'technology classification,' which involves the process of selecting relevant technology articles from Wikipedia.

Step 1: Cleaning and Selecting Relevant Categories Each Wikipedia article is linked to categories, forming a complex graph with parent-child relationships. The edges between categories are *loosely* defined as "is related to," often connecting different Wikipedia articles from non-technology areas. This correlation appears to limit the reliability of extracting only technology articles using these graph-based relationships.

To address this, we first clean up the directed categories graph by removing hidden categories, admin and user pages. Furthermore, we apply regular expression filters to eliminate categories not related to technologies, such as companies, people names, brands, currencies, and countries.

Additionally, we utilize Wikipedia's Main Topic Classifications (MTC), encompassing categories like Technology, Business, Arts, Health, etc. Subsequently, we calculate the shortest path for each category in the filtered graph corresponding to 28 MTC to retain the articles with the smallest distance to Technology, Science, or Engineering concepts. This resulted in 7,876 technology classification candidates, still containing some categories that may not belong to technology. By having a human domain expert manually go through the 7,876 technology classification candidates, we ultimately create a list of 1,356 technology categories.

Succinctly, this process can be written as the following pseudocode in Algorithm 1.

Step 2: Technology Classification using SVM The overall process of machine learning-based training to obtain the final technology classification is detailed in Algorithm 2.

To create an input dataset for the Support Vector Machine (SVM), which serves as our classifier, we extract abstracts from Wikipedia articles identified within the technology categories established in Step 1. The abstracts from all Wikipedia pages directly linked to a technology category are concatenated, stemmed, and then subjected to TF-IDF-based weighting. This process generates a weighted

[4] https://www.dbpedia.org/

Algorithm 1 Cleaning and Selecting Relevant Categories

procedure CLEANUPDIRECTEDGRAPH
    Remove hidden categories, admin and user pages from the directed categories graph
    Apply regular expression filters to eliminate irrelevant categories (e.g., companies, people names, brands, currencies, and countries)
end procedure
procedure UTILIZEMAINTOPICCLASSIFICATIONS
    Use Main Topic Classifications (MTC) encompassing categories like Technology, Business, Arts, Health, etc.
    Calculate the shortest path for each category in the filtered graph to MTC
end procedure
procedure FILTERBYDISTANCETOMTC
    Retain articles with the smallest distance to Technology, Science, or Engineering concepts within MTC
end procedure

bag-of-words for each technology category. Subsequently, feature reduction is applied to form usable feature vectors. It is worth noting that optimal results were observed using mutual information-based feature reduction, targeting a vector length of 1000. Distances to each MTC topic are appended to this vector, producing the final feature vectors as input features.

To address the imbalance in class distribution caused by our small training set of 1,356 positive samples, we employ oversampling techniques, using Borderline-SMOTE [29], to increase the size of the input samples. The list of technologies identified through SVM training is considered the final list pertaining to technology.

This final list is subsequently used to filter annotations from patents and publications.

Algorithm 2 Technology Classification using SVM

procedure CREATEDATASET
    Extract abstracts from Wikipedia articles in identified technology categories
    Concatenate and stem abstracts, apply TF-IDF-based weighting
    Perform feature reduction for usable feature vectors
    Append distances to each MTC topic to create final feature vectors
end procedure
procedure HANDLECLASSIMBALANCE
    Employ Borderline-SMOTE for oversampling
end procedure
procedure FINALIZETECHNOLOGYLIST
    Use SVM training outcome as the final list of technologies
end procedure

4.2. Emergence Score

Novelty Score: Novelty in emerging technologies signifies their distinctive newness, pioneering concepts, breakthrough advancements, and creative problem-solving, distinguishing them from existing solutions and suggesting transformative potential [15, 18].

In our study, we define novelty for a technology based on increased mentions in recent years. For instance, if a particular technology has a significant portion of references occurring in the last few years, it receives a high novelty score. To implement this, we considered the time span of the last 10 years and calculated the percentage of annotations for each year. Linearly decreasing weights ranging from 10 to 1 were assigned, respectively, thereby giving higher weight to more recent years. Technologies for which the majority of annotations occurred more than 10 years ago are considered not meeting the novelty criterion and are consequently discarded.

To express this more mathematically, we first define the yearly time series $X_{t,d}$ using Eq. 1:

$$X_{t,d} = \{X_{t,d,y} : y \in Y\} \tag{1}$$

where:

- $X_{t,d,y}$ is the number of times technology t is referenced in dataset d during year y.
- $y \in Y$ denotes the year within the specified range.

Thus, the total number of occurrences of all technologies $t \in T$ in a dataset $d \in D$ over a given year y is represented mathematically as Eq. 2:

$$\text{Total}(t, d) = \sum_{y \in Y} X_{t,d,y} \tag{2}$$

where:

- Total(t,d) denotes the total count of mentions or occurrences of technology (t) in dataset (d).
- $X_{t,d,y}$ is the number of times technology t is referenced in dataset d during year y.
- $\sum_{y \in Y}$ signifies the summation over all years (y) within the specified range Y.

The novelty score *Novelty*(t) of a technology $t \in T$ is then expressed mathematically as Eq. 3:

$$\textbf{\textit{Novelty}(t)} = \sum_{d \in D} \sum_{y \in Y} \left(\frac{X_{t,d,y}}{\text{Total}(t, d)} \times 100 \times w_y \right) \tag{3}$$

where:

- *Novelty*(t) represent novelty score for technology (t).
- $X_{t,d,y}$ is the number of times technology t is mentioned in dataset d during year y.
- Total(t,d) represents the total occurrences of technology (t) in dataset (d).
- w_y is a weight assigned to each year based on Eq. 4.
- $\sum_{d \in D} \sum_{y \in Y}$ denotes double summation over all datasets(D) and years (Y).

The formula computes the weight for each year based on its relative position within the given range. The weight increases linearly with the year's proximity to the earliest year, providing a higher weight to more recent years, as Eq. 4:

$$w_y = (y + 1 - \min_{\forall y' \in Y} y') \tag{4}$$

where:

- y denotes the specific year for which the weight is calculated.

- $\min_{\forall y' \in Y} y'$ signifies the minimum value among all years in the defined range Y.

Growth Score: Emerging technologies exhibit relatively fast growth rates compared to non-emerging technologies [18]. The growth rate of a technology, assessed through growth curves in patents and publications, has been studied extensively [30, 31, 32]. Using the concept of growth curves, we employ a two-step approach to compute the growth score of a technology.

In Step 1, we apply regression techniques to fit the number of yearly technology mentions to four different curve models: Linear, Quadratic, Gaussian, and Exponential [5]. We select the model with the highest R-squared (R_2) measure [33] and compute the slope of the curve based on the regression coefficients. It is important to note that we assume the positive or negative sign of the slope determines whether the trend is increasing or decreasing. Subsequently, based on the best-fitting model and the slope, we assign the technology to one of the classes defined in Table 1 to compute the model_score.

Table 1
Curve models and growth scores

curve model	model_score
Exponent increase/decrease	+/- 1.00
Quadratic increase/decrease	+/- 0.75
Gaussian increase/decrease	+/- 0.05
Linear increase/decrease	+/- 0.25
Nothing fits	0.00

In Step 2, the slope of the technology growth curve *Slope*(t, d) is calculated by taking the difference between the absolute counts of the last and the first year and dividing it by the total number of years, as depicted in Eq. 5. This equation quantifies the rate of change in technology mentions over time for a specific technology (t) within a dataset (d).

$$Slope(t,d) = \frac{\text{Count}(t,d,Y_{\text{final}}) - \text{Count}(t,d,Y_{\text{begin}})}{Y_{\text{final}} - Y_{\text{begin}}} \tag{5}$$

where:

- Y_{final} represents the final year for which the counts are considered.
- Y_{begin} represents the initial year for which the counts are considered.
- $Count(t, d, Y_{final})$ denotes the absolute count of mentions of the technology (t) in the dataset (d) during the final year.
- $Count(t, d, Y_{begin})$ denotes the absolute count of mentions of the technology (t) in the dataset (d) during the initial year.

Subsequently, all calculated slope values are normalized to the range [0.0;1.0] using Eq. 6, where *Norm_slope*(t, d) represents the normalized slope.

$$Norm\_slope(t,d) = \frac{Slope(t,d) - \min(Slope(T,d))}{\max(Slope(T,d)) - \min(Slope(T,d))} \tag{6}$$

[5]We utilize Apache Commons SimpleRegression and OLSMultipleLinearRegression for the linear and quadratic models. The same regression tools are used with the logarithm of the data points to derive the exponential and Gaussian models, respectively.

where:

- *Slope*(t, d) denotes the scope of the growth curve for technology (t) in dataset (d).
- $\min($*Slope*$(T, d))$ represents the minimum slope value among all technologies in dataset (d).
- $\max($*Slope*$(T, d))$ represents the maximum slope value among all technologies in dataset (d).

This normalization process facilitates comparative analysis across different technologies and datasets.

The technology's final growth score is then computed by integrating both the model score, which is determined based on the best-fitting growth curve model, and the slope score, reflecting the rate of change in the technology's mentions over time, using Eq. 7.

$$\textbf{\textit{Growth}(t)} = \sum_{d \in D} (Model\_score(t,d) + Norm\_slope(t,d)) \tag{7}$$

where:

- $Model\_score(t, d)$ denotes the model_score for the specified technology (t) in the given dataset (d).
- $Norm\_score(t, d)$ denotes the normalized slope for the specified technology (t) in the given dataset (d).
- $\sum_{d \in D}$ indicates the summation across all datasets (D) for the specified technology.

Impact Score: Wikipedia Pageviews represent the number of times a particular article has been accessed on the Wikipedia website, providing insights into the level of public interest and engagement with specific topics or content. Utilizing this information, we leverage Wikipedia Pageview statistics to compute the impact score of a technology. We use a monthly views to gather more data points. After extracting the monthly views, denoted as (w), we apply a 3-month moving average filter to smooth the time series. This filter calculates the average of each data point along with the two preceding and two succeeding months, effectively reducing noise and revealing underlying trends - see Eq. 8.

$$MA_i = \frac{w_{i-2} + w_{i-1} + w_i + w_{i+1} + w_{i+2}}{5} \tag{8}$$

The smoothed data (MA_i) then replaces (d) in the two-step approach used for the growth score. We classify the trends into the same five classes (as seen in Table 1).

$$\textbf{\textit{Impact}(t)} = Model\_score(t, MA_i) + Norm\_slope(t, MA_i) \tag{9}$$

Eq. 9 represents the calculation of the impact score *Impact*(t) for a technology (t). It combines the model score $Model\_score(t, MA_i)$ and the normalized slope score $Norm\_slope(t, MA_i)$ obtained from the 3-month moving average (MA_i) of Wikipedia Pageviews. This score reflects both the growth pattern and the temporal trends in Wikipedia Pageviews, providing a comprehensive assessment of the technology's impact.

Coherence Score: In our study, we consider coherence as the persistence of a technology over time, as referred to by [18]. When identifying emerging technologies, we assume that the presence of a category on Wikipedia signifies

a thematic grouping that brings together related technological concepts. The coherence within such categories is established through shared characteristics, applications, and underlying principles of the technologies they encompass. This alignment allows for consistent trends to emerge within the category over time, reflecting the collective evolution of technologies. Wikipedia categorization serves as a valuable indicator of how various technologies within a category develop in tandem, providing insights into the overarching trends and advancements in related technological domains.

To compute the coherence score, we begin by collecting all unique categories from Wikipedia, forming what we refer to as the 'Category Set.' Subsequently, we perform a mapping process, converting plural category names to their singular counterparts, and then matching them with articles sharing identical names. The coherence score is then computed with the following Eq. 10:

$$\textbf{\textit{Coherence}(t)} = \begin{cases} 0.5, & \text{if } t \in \text{Category Set} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

In other words, if the technology (t) is part of the Category Set, the coherence score is 0.5; otherwise, it is 0. This mathematical expression reflects the coherent presence of a technology within a specific thematic category.

Emergence Score: Towards calculating the emergence score, we sum the novelty, growth, impact, and coherence scores. We then normalize the result to the range [0.0;1.0], as shown in Eq. 11.

$$\textbf{\textit{Emergence}(t)} = Norm[n * Novelty(t) + g * Growth(t) + i * Impact(t) + c * Coherence(t)] \tag{11}$$

We introduce control variables, including n, g, i, and c, to empirically manage the impact of biases arising from data imbalance, aiming to achieve the highest precision.

Technology Class and Technology Class Score: Individuals often generate multiple articles on Wikipedia that closely relate to one another, such as those on Machine Learning, Deep Learning, and Artificial Neural Networks. To establish connections between these closely related technologies, we employ Wikidata properties such as 'subclass of,' 'part of,' 'instance of,' or 'said to be the same as.' We refer to this group of related technologies as a 'Technology Class.' The Technology Class score (TCs) is computed by taking the emergence score of the technology within the set of related technologies, selecting the one with the maximum emergence score, as shown in Eq. 12:.

$$\textbf{\textit{TCs}} = \max_{t \in EC} \text{Emergence (t)} \tag{12}$$

5. Evaluation

For patents, we gathered the abstracts of 6,647,699 patents from PatentsView. From this dataset, we derived 112,199 unique annotations, of which 77,995 had more than 5 occurrences. Similarly, for publications, we collected the abstracts of 1,425,558 research papers from arXiv. Within this dataset, we identified 111,627 unique annotations with technology classification, and among them, 65,162 articles had occurrences exceeding 5 times. Our proposed technology classification method identifies 50,954 technologies from the 4,996,310 Wikipedia articles we utilized in our study.

5.1. Results

In this section, we discuss the observations obtained after applying our proposed methodology to the public dataset discussed earlier.

Individual Scores: Table 2 displays the top 20 technologies with the highest novelty, growth, and impact scores. Notably, technologies related to Artificial Intelligence (AI)† appear among the top 20 across all scores, including Deep Learning and Convolutional Neural Network (CNN) for novelty, and Artificial Intelligence, Machine Learning, and Artificial Neural Network for impact; all except CNN correspond to categories in Wikipedia and are considered coherent.

In the top 20 novel technologies, alongside AI-related technologies, there are notable mentions of vehicle-related technologies such as Multirotor, Autonomous Car, and Vehicle-to-everything. The Nanosheet closes the novelty list, being the only technology not related to either computer science or vehicle technology. Communication ranks first in the list of the top 20 technologies according to the growth score, with Communication-related technologies like Wireless and Data Transmission being other fast-growing terms. The list also includes older technologies that receive continuous or renewed interest, such as Lidar or Rechargeable Battery. Apart from vehicle-related technologies like Unmanned Aerial Vehicle and Autonomous Car, this list is completed by the Internet of Things and Quantum Computing.

Overall Score: Table 3 presents the overall top 20 technologies after combining the individual scores.

Deep Learning emerges as the top technology in our methodology, with Convolutional Neural Network (CNN) also making the list as a sub-category of Deep Learning. As anticipated, Machine Learning is present, alongside the Internet of Things, both demonstrating coherence and ranking in the top 20 for impact and novelty, respectively. Cyberattack holds a high position, accompanied by various technologies related to Computer security, forming the second group in the result list. Key-Value Database, the simplest form of NoSQL databases, secures the seventh spot in the top 20 emerging technologies. Communication and Smartphone, technologies that have garnered attention for years, are also on the list. We observe the inclusion of technologies such as Autonomous Car, Knowledge Graph, and 5G in the top 20 scored technologies.

Our findings align well with similar observations made by Zhou et al. [34] and Daim et al. [35], returning four Convergence Emerging Technologies (CET) in the top five results, with the fifth (CNN) being a sub-class of Deep Learning.

Table 4 displays the top 20 technology classes identified from the top 100 technologies based on the emergence score. This method of presenting results enhances the visibility of other technologies, such as Virtual Assistant or Exoskeleton.

5.2. Benchmarking

To benchmark the compatibility of our proposed emergence scoring to other similar works, we compiled the union set of emerging technologies identified by leading technology analysts, including Gartner, Forrester, IHS Markit, and

Table 2
Top 20 Technologies in Novelty, Growth, and Impact scores

Novelty	Growth	Impact
Smart City	Communication	URL
Deep Learning†	Wireless	LED Lamp
POWER8	Pixel	Machine Learning†
Vehicle To Everything	Web Server	Artificial Neural Network†
Data Science	Convolutional Neural Network†	Neural Coding
Knowledge Graph	Data Transmission	Robot Locomotion
Internet of Things	Mathematical Optimization	HTTP Cookie
Return-Oriented Programming	Stator	Blockchain
Smartwatch	Rechargeable Battery	Artificial Intelligence†
Multirotor	Radio-Frequency Identification	Computer Science
Ransomware	Unmanned Aerial Vehicle	Sustainable Energy
Row Hammer	Internet of things	BNC Connector
Software-Defined Networking	Quantum Computing	Electron Backscatter Diffraction
Convolutional Neural Network†	Computer Data Storage	Slurry Pump
Virtual Reality Headset	Object Detection	Cryptocurrency
High Efficiency	Video Coding	Lidar Precision and Recall
Cyber-Physical System	Transfer Learning†	XLR Connector
Insider Threat	Unsupervised Learning†	Phishing
Autonomous Car	HVAC	QR Code
Nanosheet	Autonomous Car	PDF

Table 3
Overall Top 20 Technologies

Technology
Deep Learning†
Autonomous Car
Internet of Things
Convolutional Neural Network (CNN)†
Machine Learning†
Ransomware
Key-Value Database
Shard (Database Architecture)
Cyberattack
Knowledge Graph
Augmented Reality
Smartphone
Communication
Side-Channel Attack
Cloud Gaming
5G
Data Science
Return Oriented Programming
Lidar
Push Technology

Table 4
Overall Top 20 Technology Classes

Technology Classes
Artificial Intelligence
Autonomous Driving
Internet of Thing
Computer Security
Database
Knowledge Graph
Augmented, Virtual, Mixed Reality
Connectivity
Telecommunication
Cloud and Virtualization
Data Science
Optical Instrument
Virtual Assistant
Exoskeleton
Computer Vision
Satellite Imagery
Heterogeneous Computing
Distributed Computing
Medical Device
3D Printing

the World Economic Forum (WEF). Gartner predicted 35 technologies in its technology hype cycle, Forrester predicted 12, IHS Markit 8, and WEF 10 emerging technologies. Upon merging the overlapping technologies from these four lists, we derived a consolidated list of 36 unique technology classes which we use as ground truth. Table 5 provides an overview of these classes.

Notably, the majority of technologies in this table appear to belong to the Computer Science-related domain, with 72% of them being linked to it. Technologies marked with '†' are those we were unable to directly map to a Wikipedia article or category. Additionally, articles judged as non-technologies by the SVM classifier are indicated in the table with '.'

It is worth mentioning that Wikipedia articles on Augmented, Mixed, and Virtual Reality are collectively presented, following Forrester's proposal to consider them as a single technology class.

Table 6 illustrates the performance metrics of Average Precision (AP) and Recall (R) for the top 20 technologies (T) and Technology Classes (TC) identified in the evaluation set.

In the 'base' run, all control variables in Eq. 10 are set to 1. Additionally, alongside the 'max_prec' parameter set, we present the average precision and recall of the Computer Science technology class (max_prec_cs). Within the top 20 technologies with the highest emergence score, only one non-technology result was observed. The average precision (AP) was 0.72 for the base run. However, all the relevant concepts from this subset relate to only 6 out of the 36 technologies mentioned before, resulting in a recall (R) of 0.16. By changing the control variables for the max_prec, where non-Computer Science technology does not grow and have entries in Wikipedia articles, we were able to increase

Table 5
Evaluation Set: Technology classes based on Gartner, Forrester, IHS Markit and WEF

Technology Classes
Tissue Engineering
Unmanned Aerial Vehicle
Smartdust
Artificial Intelligence
4D Printing
Ontology (Information Science)
Neuromorphic Engineering
Exoskeleton
Edge Computing
Autonomous Driving
Self-Healing System Technology†
Volumetric Display
5G
Quantum Computing
Platform as a Service
Application Specific Integrated Circuits
Autonomous Robot
Mobile Robot
Brain Computer Interface
Internet of Things
Biochip
Digital Twin
Nanotechnology
Virtual Assistant
Lithium-Silicon Battery
Blockchain
Augmented, Virtual, Mixed Reality
E-textiles
Cloud Computing
Computer Vision
Ubiquitous Video†
Natural Language Generation
Switched Fabric
Personalized Medicine
Cell Encapsulation
Gene drive

Table 6
Average Precision (AP) and Recall (R) of Technologies (T) and Technology Classes (TC)

Parameters	Classes	AP	R
base	T	0.72	0.16
max_prec	T	0.81	0.19
	TC	0.72	0.28
	CS TC	0.79	0.36
max_prec_cs	CS TC	0.90	0.36

both AP (0.81) and R (0.19). In this setting, the control variables were chosen to facilitate the maximum precision (e.g., g, n, i, and c set to 1, 0.3, 0.1, and 0.3, respectively).

6. Limitations

A bias is evident when examining the results of identified emerging technologies toward Computer Science, as noticed within the evaluation set, with 70% of technologies within the top 100 results belonging to this domain. This bias complicates the exploration of trends in other domains. Taking chemistry as an example, the International Union of Pure and Applied Chemistry (IUPAC) issued a list of emerging technologies for this domain, containing, among others, 3D bioprinting or Flow chemistry, none of which figure in our evaluation set but are present in our technology result set, ranked 4,897 and 12,421, respectively. To address this bias, we split the result set as well as the evaluation set into distinct domains (CS, Nanotechnology, Medicine, etc.). This approach allowed us to navigate around the bias. The third row (CS TC) of Table 6 provides the average precision and recall when only results related to the Computer Science field are considered, as this class is predominant in our result/evaluation sets. Although this approach results in only a 10% increase in average precision, the increase in recall rises to 30%.

7. Conclusion

This paper presents an automated method for identifying emerging technologies using publicly available data. Our approach is applicable across various technology sectors without the need for human domain experts, as it relies on a clear mathematical foundation.

We propose an emergence scoring system based on novelty, growth, impact, and coherence scores. Novelty and growth scores are computed from time series data of annotations applied to USPO patents and arXiv publications. The impact score is derived from the Wikipedia Pageview time series, while the coherence score utilizes Wikipedia categories.

To assess the effectiveness of our proposed methods, we compiled an evaluation set of 36 emerging technologies by amalgamating lists from prominent market research firms like Gartner and Forrester Research. The evaluation unveiled a low recall (0.16) in identifying emerging technologies.

This research lays the groundwork for further investigations, including the development of a methodology to determine the more fine-grained stages of emergence (e.g., pre-emergence, emergence, post-emergence) for a particular technology within different timeframes.

Our study can be enhanced by incorporating the OpenAlex concept [6], which has gained more popularity compared to the now-defunct DBpedia concepts. Additionally, we plan to employ more advanced deep learning models instead of the SVM model, as mentioned in [36, 37], specifically a combination of LSTM and Transformer [38, 39], to conduct more efficient time series analysis. This will be performed using a larger publication dataset than arXiv, such as the one available on OpenAlex [7]. Additionally, since our methodology still requires a certain degree of manual intervention, such as inspecting Wikipedia categories and adjusting bias variables, we want to explore techniques that can minimize these manual components to enhance scalability and reduce potential subjectivity.

Acknowledgments

We extend our thanks to the developers at Trivo Systems—Pratiksha Jain, Himanshu Jain, and Marc Liechti—for their work on the Technology Market Monitoring 1.0 project. We appreciate their valuable contributions to shaping the initial stage of our study. We also extend our thanks to armasuisse Science and Technology for supporting the study.

[6] https://docs.openalex.org/api-entities/concepts
[7] https://openalex.org/

References

[1] O. Dedehayir, M. Steinert, The hype cycle model: A review and future directions, Technological Forecasting and Social Change 108 (2016) 28–41.

[2] G. Intepe, T. Koc, The use of s curves in technology forecasting and its application on 3d tv technology, International Journal of Industrial and Manufacturing Engineering 6 (2012) 2491–2495.

[3] S. Ranaei, M. Karvonen, A. Suominen, T. Kässi, Forecasting emerging technologies of low emission vehicle, in: Proceedings of PICMET'14 Conference: Portland International Center for Management of Engineering and Technology; Infrastructure and Service Integration, IEEE, 2014, pp. 2924–2937.

[4] J. W. Z. Sossa, F. P. Marro, B. A. Alzate, F. M. V. Salazar, A. F. A. Patiño, S-curve analysis and technology life cycle. application in series of data of articles and patents, Revista ESPACIOS| Vol. 37 (N° 07) Año 2016 (2016).

[5] S. Kar, A. K. Kar, M. P. Gupta, Understanding the s-curve of ambidextrous behavior in learning emerging digital technologies, IEEE Engineering Management Review 49 (2021) 76–98.

[6] R. Adner, R. Kapoor, Innovation ecosystems and the pace of substitution: Re-examining technology s-curves, Strategic management journal 37 (2016) 625–648.

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

[8] B. R. Martin, Foresight in science and technology, Technology analysis & strategic management 7 (1995) 139–168.

[9] N. Corrocher, F. Malerba, F. Montobbio, The emergence of new technologies in the ICT field: main actors, geographical distribution and knowledge sources, Technical Report, Department of Economics, University of Insubria, 2003.

[10] M. Halaweh, Emerging technology: What is it, Journal of technology management & innovation 8 (2013) 108–115.

[11] S.-C. Hung, Y.-Y. Chu, Stimulating new industries from emerging technologies: challenges for the public sector, Technovation 26 (2006) 104–110.

[12] W. Boon, E. Moors, Exploring emerging technologies using metaphors–a study of orphan drugs and pharmacogenomics, Social science & medicine 66 (2008) 1915–1927.

[13] S. Cozzens, S. Gatchair, J. Kang, K.-S. Kim, H. J. Lee, G. Ordóñez, A. Porter, Emerging technologies: quantitative identification and measurement, Technology Analysis & Strategic Management 22 (2010) 361–376.

[14] B. C. Stahl, What does the future hold? a critical view of emerging information and communication technologies and their social consequences, in: Researching the Future in Information Systems: IFIP WG 8.2 Working Conference, Turku, Finland, June 6-8, 2011. Proceedings, Springer, 2011, pp. 59–76.

[15] H. Small, K. W. Boyack, R. Klavans, Identifying emerging topics in science and technology, Research policy 43 (2014) 1450–1467.

[16] W. Glänzel, B. Thijs, Using 'core documents' for detecting and labelling new emerging topics, Scientometrics 91 (2012) 399–416.

[17] A. Tavazzi, D. P. David, J. Jang-Jaccard, A. Mermoud, Measuring technological convergence in encryption technologies with proximity indices: A text mining and bibliometric analysis using openalex, arXiv preprint arXiv:2403.01601 (2024).

[18] D. Rotolo, D. Hicks, B. R. Martin, What is an emerging technology?, Research policy 44 (2015) 1827–1843.

[19] T. U. Daim, G. Rueda, H. Martin, P. Gerdsri, Forecasting emerging technologies: Use of bibliometrics and patent analysis, Technological forecasting and social change 73 (2006) 981–1012.

[20] D. Kucharavy, E. Schenk, R. De Guio, Long-run forecasting of emerging technologies with logistic models and growth of knowledge, in: 19th CIRP design conference, 2009, p. 277.

[21] M. Bengisu, R. Nekhili, Forecasting emerging technologies with the aid of science and technology databases, Technological Forecasting and Social Change 73 (2006) 835–844.

[22] M. Nieto, F. Lopéz, F. Cruz, Performance analysis of technology using the s curve model: the case of digital signal processing (dsp) technologies, Technovation 18 (1998) 439–457.

[23] M. N. Kyebambe, G. Cheng, Y. Huang, C. He, Z. Zhang, Forecasting emerging technologies: A supervised learning approach through patent analysis, Technological Forecasting and Social Change 125 (2017) 236–244.

[24] S.-Y. Hwang, D.-J. Shin, J.-J. Kim, Systematic review on identification and prediction of deep learning-based cyber security technology and convergence fields, Symmetry 14 (2022) 683.

[25] Y. Zhou, F. Dong, Z. Li, J. Du, Y. Liu, L. Zhang, Forecasting emerging technologies with deep learning and data augmentation: convergence emerging technologies vs non-convergence emerging technologies (2017).

[26] P. USPTO, Locations that drive innovation, 2023. URL: https://datatool.patentsview.org/, accessed: December 9, 2023.

[27] arXiv, Monthly submissions, 2024. URL: https://arxiv.org/stats/monthly_submissions, accessed: February 5, 2024.

[28] P. Analysis, Comparison of pageviews across multiple pages, 2023. URL: https://pageviews.wmcloud.org/, accessed: February 12, 2024.

[29] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-smote: a new over-sampling method in imbalanced data sets learning, in: International conference on intelligent computing, Springer, 2005, pp. 878–887.

[30] B. Andersen, The hunt for s-shaped growth paths in technological innovation: a patent study, Journal of evolutionary economics 9 (1999) 487–526.

[31] M. Meyer, Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology, Scientometrics 51 (2001) 163–183.

[32] G. S. Day, P. J. Schoemaker, Avoiding the pitfalls of emerging technologies, California management review 42 (2000) 8–33.

[33] D. S. Moore, Introduction to the Practice of Statistics, WH Freeman and company, 2009.

[34] Y. Zhou, F. Dong, Y. Liu, Z. Li, J. Du, L. Zhang, Forecasting emerging technologies using data augmentation and deep learning, Scientometrics 123 (2020) 1–29.

[35] T. Daim, K. K. Lai, H. Yalcin, F. Alsoubie, V. Kumar, Forecasting technological positioning through technology knowledge redundancy: Patent citation analysis of iot, cybersecurity, and blockchain, Technological

Forecasting and Social Change 161 (2020) 120329.
[36] Y. Zhang, C. Zhang, P. Mayr, A. Suominen, Y. Ding, An editorial of "ai+ informetrics": Robust models for large-scale analytics, Information Processing and Management (2023) 103495.
[37] W. Xu, J. Jang-Jaccard, A. Singh, Y. Wei, F. Sabrina, Improving performance of autoencoder-based network anomaly detection on nsl-kdd dataset, IEEE Access 9 (2021) 140136–140146.
[38] Y. Wei, J. Jang-Jaccard, W. Xu, F. Sabrina, S. Camtepe, M. Boulic, Lstm-autoencoder-based anomaly detection for indoor air quality time-series data, IEEE Sensors Journal 23 (2023) 3787–3800.
[39] Y. Wei, J. Jang-Jaccard, F. Sabrina, W. Xu, S. Camtepe, A. Dunmore, Reconstruction-based lstm-autoencoder for anomaly-based ddos attack detection over multivariate time-series data, arXiv preprint arXiv:2305.09475 (2023).