Early Detection of Emerging Technologies Using Machine Learning and Burst Detection

Ali Ghaemmaghami

A Thesis
In the Department of
Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Applied Science in Quality Systems Engineering
at Concordia University
Montréal, Quebec, Canada

November 2024

CONCORDIA UNIVERSITY SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: *Ali Ghaemmaghami*

Entitled: *Early Detection of Emerging Technologies Using Machine Learning and Burst Detection*

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Quality Systems Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Yong Zeng

_____ Examiner
Dr. Yong Zeng

_____ Examiner
Dr. Arash Mohammadi

_____ Thesis Supervisor
Dr. Andrea Schiffauerova

_____ Thesis co-supervisor
Dr. Ashkan Ebadi

Approved by

_____
Dr. Farnoosh Naderkhani, Graduate Program Director
Department of Concordia Institute for Information Systems Engineering

Year 2024

_____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

ABSTRACT

Early Detection of Emerging Technologies Using Machine Learning and Burst Detection

Ali Ghaemmaghami

Certainly, the impact of emerging technologies is changing our world and how we live, shaping our future significantly. In the constantly evolving landscape of these technologies, which attracts substantial yearly investments, spotting these trends early on is both challenging and expensive. However, applying an emerging technology detection method in an effective and efficient way is considered a challenging task for many stakeholders. In this thesis, we address these problems through applying a method to predict potential emerging technologies in the case study field of Artificial Intelligence (AI). Using this method may help policymakers to identify potential emerging technologies early in a more systematic way with little manual intervention. In the proposed method, using burst detection, machine learning, and deep learning, we attempt to predict the future sustaining emerging technologies. We applied the methodology by four methods, namely Random Forest, Gradient Boosting, XGBoost, and Multi-Layer Perceptron (MLP). Results showed that the method was successful in its tasks. The method had the Area under the Curve (AUC) rate of more than 75% to accurately predict the sustainability of the potential emerging technologies. More specifically, applying the MLP method showed the ability to increase the AUC rate and recall metric as the most important metrics of our work. In summary, this approach carries both theoretical and practical significance. Theoretically, the exploration of novel combinations, such as integrating deep learning and burst detection methods or employing transformers, offers researchers fresh insights into the challenge of detecting emergence. On the practical front, the application of methods providing high accuracy rates in machine learning methods empowers stakeholders to implement these methods effectively in practical scenarios.

ACKNOWLEDGMENT

I am immensely grateful to all those who have contributed to the completion of this Master's thesis, and I would like to take this opportunity to express my heartfelt appreciation to the following individuals:

First and foremost, I express my deepest gratitude to my supervisor Dr. Ashkan Ebadi and Professor Andrea Schiffauerova. Their unwavering support, guidance, and encouragement throughout this research journey have been invaluable. Their expertise, insightful feedback, and constructive criticism have played a pivotal role in shaping this thesis and refining my understanding of the subject matter.

To my loving parents, I owe an immeasurable debt of gratitude. Your belief in my abilities and constant encouragement have been the driving force behind my academic pursuits. Your sacrifices, both emotional and financial, have made it possible for me to pursue higher education and reach this milestone in my life. I am eternally thankful for your unconditional love, support, and the values you have instilled in me.

I would also like to extend my heartfelt thanks to my dear wife, Saeide. Your unwavering love, patience, and understanding have been the anchor that kept me steady during the challenging times of this thesis. Your constant encouragement, listening ear, and belief in my abilities have been a source of inspiration. Your support throughout this journey have meant the world to me, and I am forever grateful to have you as my partner in life.

Lastly, I want to acknowledge the support and understanding of my friends, who have cheered me on and provided encouragement and motivation along the way. Your belief in me has been a great source of strength.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and Motivation

Emerging technologies have the potential not only to alter the technological paradigms on which traditional industries rely and to generate entirely new industries (Day and Schoemaker 2000; Porter et al. 2002), but also to alter existing socio-economic structures and production practices (Adner and Snow 2010; Rotolo et al. 2015; Y. Zhou et al. 2019). Early and accurate detection of emerging technologies can provide decision-makers with knowledge, intelligence, and opportunities, from research and development (R&D) departments of different institutions to national policy-making organizations and innovation administrations (Jang et al. 2021; S. Xu et al. 2021; Y. Zhou et al. 2021). Especially, in more recent years, the speed at which technology changes and advances has been staggering, making fast detection of emerging technologies more valuable. For instance, popular cutting-edge technologies such as cloud computing, mobile computing, the Internet of Things, the Internet of Services, data collection, big data analytics, artificial intelligence, augmented reality and 3D printing developed and adopted quickly (Zamani et al. 2022). As the amount of data available to us is growing at a stunning pace, applications of this data are growing as well. One of the applications of this huge quantity of data can be detecting emerging technologies or topics through data and with minimal intervention of experts.

The trend on the topic of emerging technologies has followed an upward trajectory in recent years. The trend regarding results for the defined search with the topic of “emerging technologies” in the Web of Science (WOS) database since the year 2000 is represented in Figure 1. The topic of emerging technologies has been considered interesting in recent years, and thereby the subtopics around it, such as emerging technologies detection, have similarly gained attention.

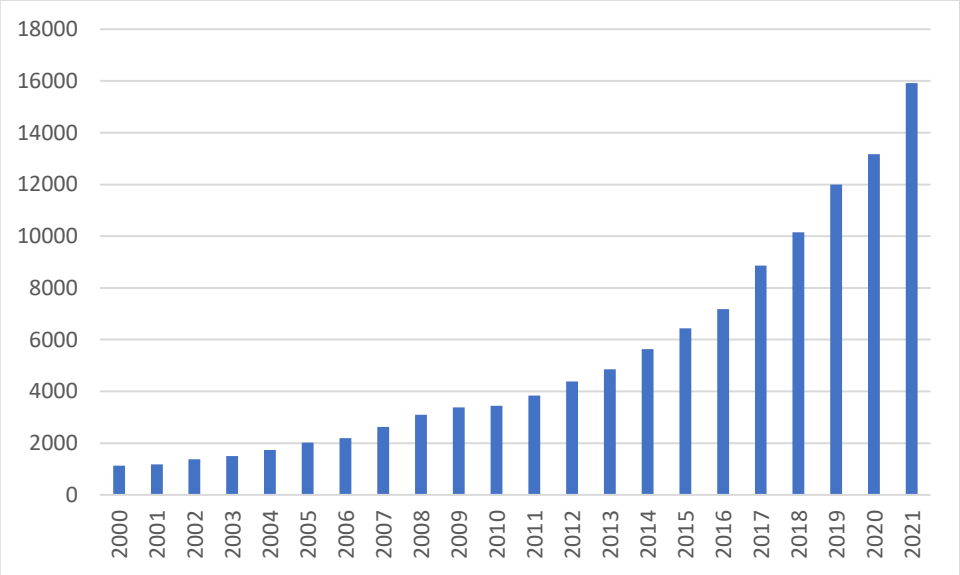

Figure 1. The trend of the number of documents with the topic of emerging technologies in the WOS database

There are different definitions of emerging technologies. Many of these definitions rely on attributes of emerging technologies, such as the famous definition of Rotolo et al. (2015) that defined emerging technologies with five attributes: radical novelty, relatively fast growth, coherence, prominent impact as well as uncertainty and ambiguity.

Over the years, manifold approaches to emergence detection have been deployed to identify emerging technologies, from lexical-based approaches (Joung and Kim 2017; Weismayer and Pezenka 2017; Wu and Leu 2014), bibliometric approaches (Daim et al. 2006; Kim and Bae 2017; Mejia and Kajikawa 2020), and indicator-based approaches (Abercrombie et al. 2012; Bengisu 2003; H. Xu et al. 2021) to more complex methods such as machine learning methods (Choi et al. 2021; S. Xu et al. 2021; Y. Zhou et al. 2021) and hybrid methods (Ávila-Robinson and Miyazaki 2013; Carley et al. 2017; Wang 2018). A new weak signal analysis method has been added to the emergence detection methods in recent years (Ebadi et al. 2022).

Most of these approaches use patents or publications as their source of emerging technologies. Based on their approach and their focus on various aspects, researchers tried to choose between patents and papers as their source of data. Both patents and papers can provide information about emerging technologies, but at various times and levels. Ávila-Robinson and Miyazaki (2013) deployed both sources to capture the cycle

of emerging technologies. However, very few tried to include both in the process of emerging technology detection. Mejia and Kajikawa (2020) were one of the few that evaluated the emerging topics in both science and technology using paper and patent databases simultaneously. Deploying both sources in a method can yield the opportunity of using the maximal quantity of all the useful data to meaningfully detect emerging technologies.

Many of the detection methods rely on emergence attributes, but there is no consensus regarding these attributes. Rotolo et al. (2015) considered the five aforementioned features as attributes of emergence. Wang (2018) took novelty, relatively fast growth, coherence, and scientific impact as attributes of emerging research topics. Carley et al. (2017) used novelty, growth, persistence, and community as attributes of emergence in their method. However, the methods can further be refined through the addition of new perspectives of emergence to assess different dimensions of emerging technologies.

The process of detecting emerging technologies or topics does not end with the detection. Various institutions or companies want to know not only which technologies are emerging, but also which ones have higher probability of being emerging. Measuring emergence potential can be another crucial step in the process of emerging technology detection.

However, there are some limitations that have been observed in traditional approaches to detecting emerging technologies: 1) subjectivity risk due to manual interventions, 2) lack of scalability, 3) lack of quantifiable metrics to determine the performance of the emerging technology detection process, and 4) lack of predictability or low accuracy rates for future predictions.

In this thesis, we are going to introduce and employ an emerging technology detection method in which the emerging technology terms in the field of AI will be detected using a burst detection algorithm. Then, the future trends of these terms will be predicted by combining machine learning techniques with burst detection, producing well-defined quantitative metrics for performance evaluation. The burst detection method is automated which reduces the need for manual curation and the risk of human biases.

The structure of this thesis is outlined as follows: the next section outlines the research objectives; Chapter 2 provides a review of pertinent research; Chapter 3 delves into the data utilized in this study; Chapter 4 presents the methodology of the thesis; Chapter 5 represents the results for the research objective; Chapter 6 discusses the findings of this research; and Chapter 7 concludes the thesis by highlighting the limitations of this study while suggesting directions for future research.

1.2 Research Objectives

To address the limitations mentioned in the previous section, the main objective of this research is as follows:

Objective: Prediction of sustaining emerging technologies using burst detection and machine learning

- Proposing an emerging technology term detection framework that requires little tuning and manual interventions,
- Testing the proposed approach on a case technology field, i.e., artificial intelligence (AI), using different datasets.

2 Literature Review

This section provides a concise summary of the literature relevant to the research objective. It begins by examining studies that scrutinize the broader research landscape of emerging technology detection, comparing findings in previous research. Subsequently, it delves into literature that explores burst detection in emerging technologies.

Identifying the most suitable method for future implementation is crucial. To achieve this, we need to comprehend the key methods for detecting emergence. Furthermore, because various data types have been used and proposed in the literature, we need to understand the research gaps in order to deduce implications for the data types of emergence detection as well.

2.1 Definitions of Emerging Technologies

There are various definitions for the concept of emerging technologies, and, depending on the definition, different methods for detecting emerging technologies are applicable. The main focus of this research is on the following definitions that utilize defined attributes to detect emerging technologies. Cozzens et al. (2010) defined emerging technologies as those with characteristics such as rapid growth, newness, untapped market potential, and a high-technology base. Rotolo et al. (2015) believed that there are five attributes for considering a concept to be an emergence of novel technology: radical novelty, relatively fast growth, coherence, prominent impact, and uncertainty and ambiguity. Based on different circumstances, these definitions can be altered; for instance, Wang (2018) defined emergence for research topics by replacing prominent impact by scientific impact. Based on these definitions, many articles have been written to design approaches to detect emerging technologies.

This study does not go further about definitions of emerging technologies, firstly because previous works have dedicated a significant amount of effort to defining emerging technologies and topics; for example, the paper of Rotolo et al. (2015) discovered many of these definitions of emerging technologies in the literature. Secondly, this study does not focus on this matter, as researchers can modify the definition of emerging technologies that has been proposed by Rotolo et al. (2015) based on different circumstances. The focus of this study is on approaches to detect emerging

technologies and how to modify and improve current methods, as well as to suggest novel approaches.

2.2 Approaches to Detect Emerging Technologies

Thus far, several approaches have been proposed for detecting emerging technologies. Rotolo et al. (2015) grouped them into five classes: 1. Indicators and trend analysis; 2. Citation analysis; 3. Co-word analysis; 4. Overlay mapping; and 5. Hybrid approaches. Also, Xu et al. (2021) grouped methods of emergence detection into three groups: citation-based approaches, lexical-based approaches, and machine learning approaches. This section provides an updated categorization reflecting the latest approaches.

According to the adopted methodology, emerging detection approaches can be organized into six distinct groups: 1. Lexical-based approaches; 2. Bibliometric-based approaches; 3. Indicator-based approaches; 4. Machine learning approaches; 5. Weak signal analysis; and 6. Hybrid approaches. By classifying and examining these diverse approaches, a comprehensive understanding of the current landscape in emerging technology detection can be achieved.

2.2.1 Lexical Approaches

Lexical methods in detecting emerging technologies refer to the use of term-related information to analyze and extract information from text-based sources, such as scientific literature, patents, and news articles (Xu et al. 2021). These techniques are used to identify and track new or developing technologies by analyzing the language and terms used to describe them. Lexical methods can include techniques such as co-word analysis, and keyword analysis. These methods can be used to identify key terms and phrases associated with specific technologies, to classify documents by technology, and to identify patterns and trends between terms in research and development.

Wu and Leu (2014) recommend using a patent co-word map analysis (PCMA) in order to assess the tendencies of technological trends in the field of hydrogen energy. Furukawa et al. (2015) propose a method to analyze chronological changes in research themes as seen from proceedings articles and conference sessions in order to discover,

identify, and analyze the evolutionary process of new technologies in the numerous rapidly expanding research domains. Joung and Kim (2017) apply a keyword-based model in contents-based patent analysis and suggest a technical keyword-based analysis of patents to track developing technologies.

Weismayer and Pezenka (2017) offer a longitudinal latent semantic analysis of keywords as an application to content analytics.

Recently, lexical-based approaches have mostly been replaced with natural language processing methods that are more complex and comprehensive than lexical approaches. These new methods are included in the machine learning methods.

In summary, lexical methods can provide valuable insights into the attention and impact of emerging technologies based on the key terms containing them, but it should be used in conjunction with other methods and approaches to get a more complete understanding of the field and its trends, and to consider other important aspects of emerging technologies.

2.2.2 Bibliometric Approaches

Bibliometric methods in detecting emerging technologies refer to the use of various metrics and techniques to analyze and extract information from scientific literature, patents, and other text-based sources (Rotolo et al. 2015). Bibliometric methods are used to identify patterns and trends in research and development, and to track the evolution of specific technologies over time. These methods can include techniques such as citation analysis, co-citation analysis, and social network analysis. Bibliometric methods can be used to identify key researchers, organizations, and institutions in a specific technology area, to classify documents by technology, and to identify patterns and trends in research and development.

Daim et al. (2006) were some of the first authors to use this approach by combining the use of bibliometrics and patent analysis with well-known technology forecasting methods including scenario planning, growth curves, and analogies for three emerging technological sectors.

Shibata et al. (2009) performed three citation network methods to detect a research front including co-citation, direct citation, and bibliographic coupling, and they found that direct citation performs the best to detect large and young clusters earlier. Shibata et al.

(2008) successfully used topological measures for detecting branching innovation in the citation network of scientific publications. Shibata et al. (2011) detected emerging research fronts and future core papers by using the topological clustering method and citation network analysis. Chen et al. (2011) deployed the logistic growth curve approach to detect emergence, growth, maturity, and saturation in the field of hydrogen energy using patents.

Iwami et al. (2014) identified emerging leading papers using time transition of centrality measures. Yoon and Kim (2012) used outlier patents as sources of emerging technologies and semantic patent analysis as sources of topics. Kim and Bae (2017) formed technology clusters and with the usage of patent indicators assessed whether or not a technology cluster is promising or not.

In order to suggest potential future research topics for technological observatories, Santa Soriano et al. (2018) examined citation patterns and co-occurrence keywords while evaluating their importance and level of maturity. Mejia and Kajikawa (2020) used a computational algorithm based on citation networks and thoroughly examined energy storage emerging topics by mining journal articles and patents.

It is generally accepted that bibliometric methods can provide valuable insights into emerging technologies' attention and impact. However, in order to obtain a full understanding of the field and its trends, and to consider other important aspects of emerging technologies, a combination of this method and other approaches may be employed.

2.2.3 Indicator Approaches

Indicator methods in detecting emerging technologies refer to the use of various metrics or indicators to identify and track new or developing technologies. These indicators can be based on various data sources, such as scientific literature, patents, news articles, and social media, and can be used to identify patterns and trends in research and development.

Porter and Detampel (1995) were some of the first researchers that used the number of records that include a specific keyword in their abstract as an indicator of emerging technologies to detect in bibliometric databases. Watts and Porter (1997) introduced five indicators of Research and Development (R&D) profiles: the number of items in databases such as Science Citation Index as Fundamental research, the number of items

in databases such as Engineering Index as Applied research, the number of items in databases such as U.S. Patents as Application, the number of items in databases such as Newspapers Abstracts Daily as Fundamental research, issues raised in the Business and Popular Press Abstracts as Societal Impacts. Bengisu (2003) used the slope of the regression line of the number of records in the specific field to the time as an indicator of emergence in fields. Watts and Porter (2003) defined some cluster quality measures including cohesion (as the cosine similarity measure), entropy, and F-measure as emergence indicators. Bettencourt et al. (2008) used an epidemic model to relate the increasing number of publications and new authors in an emerging field. Schiebel et al. (2010) used interesting indicators such as TF-IDF (Term Frequency Inverse Document Frequency) and Gini coefficient as well as the minimum number of articles that contain the keyword to detect an emerging research issue. This research can be considered as one of the first to detect emerging topics with multi-layer filtering approaches and indicators. Guo et al. (2011) deployed three indicators of the number and type of bursting terms, the number of new authors in a field, and the interdisciplinary of paper references to identify emerging research areas. Järvenpää et al. (2011) used Technology Life Cycle indicators based on the databases that have been used in emerging technologies detection including the number of articles in the science datasets, the number of patents in the patent datasets, and the amount of news in newspaper datasets. Abercrombie et al. (2012) constructed a network of scholarly publications, citations, patents, news, and online mappings to discover the relations of the indicators of each, for any emerging technology. Jun (2012) evaluated the technology hype cycle of hybrid cars and used Google search traffic trend (or Google trend) as an indicator of users' behavior. Jun et al. (2014) believed that Google search trend can be a better measurement of new technology adoption than other indices such as patents, news, or articles for forecasting demands. De Rassenfosse et al. (2013) believed that the inventor's total number of priority patent applications, regardless of the patent office where they were submitted, can be an indicator to assess and detect emerging technologies with. Ho et al. (2014) used a fitted logistic curve on the cumulative number of publications in a field per year to assess the emergence and predict the life cycle of that technology.

H. Xu et al. (2021) used eight indicators to detect emergence, to wit average growth rates of paper numbers, journal numbers, funding numbers, authors numbers, weakly

connected components, strongly connected components, plus publications cited by patents divided by the total number of publications on a topic, and patents cited by publications divided by the total number of publications on a topic.

In general, indicator-based approaches can provide valuable insight into the attention and impact of emerging technologies. However, they should also be used along with other methods and approaches to gain a deeper understanding of the field and its trends, as well as to take into account other important aspects of emerging technologies.

2.2.4 Machine Learning Approaches

Machine learning methods are used to detect and track emerging technologies (Xu et al. 2021). This can include using data mining and statistical analysis to identify patterns and trends in research and development, natural language processing to analyze scientific literature and patents, and predictive modeling to forecast future technological advancements. The goal of using machine learning in emerging technology detection is to automate the process of identifying and analyzing new technologies, and to provide insights that can aid in the development and prediction of these technologies.

Because of the successful usage of machine learning approaches in different fields, the implementation of machine learning approaches has increased in the field of emerging technology detection as well. S. Xu et al. (2019) used the Dynamic Influence Model (DIM) to detect topics and then by using Citation Influence Model (CIM), they calculated input indicators and predicted the next two years' values with Multi-Task Least-Squares Support Vector Machine (MTLS-SVM). Zhou et al. (2019) combined a semi-supervised text-clustering model (Labeled Dirichlet Multi Mixture) for topic segmentation and a sentence-level semantic description method (Various-aspects Sentence-level Description) information extraction method for topic description to identify emerging technologies using a semi-supervised topic clustering model. Zhou et al. (2020) built a supervised machine learning model to label them ET (Emerging Technology) or NET (Not Emerging Technology) and patent features as inputs with the usage of data augmentation with GAN (Generative Adversarial Networks) to build enough data to train the model. Altuntas et al. (2020) evaluated emerging candidates with the patent analysis using a semi-supervised clustering. Ma et al. (2021) proposed a hybrid

approach to integrate topic modeling, semantic SAO analysis, machine learning, and expert judgment, identifying technological topics and potential development opportunities. Zhou et al. (2021) deployed 11 patent indicators to detect emerging technologies that are large-scale outlier patents using technological and social impact with the deep learning method. Jang et al. (2021) used expert opinions on future and emerging technologies identified through the LDA (Latent Dirichlet Allocation) model and fuzzy c-means probabilistic clustering by utilizing diversity and centrality indices. Choi et al. (2021) deployed 3 semi-supervised active learning algorithms with 32 input variables of patents and one binary target variable of being promising or not to identify emerging promising technologies.

Overall, machine learning can be a powerful tool for detecting emerging technologies, but it is important to recognize its limitations to implement the model and analysis of the results.

2.2.5 Weak Signal Analysis

Weak signal analysis methods in detecting emerging technologies refer to techniques used to identify and track new or developing technologies by analyzing early signs or indications of their emergence (Ebadi et al. 2022). These methods are used to identify potential future technologies by detecting patterns and trends in data that may not be immediately apparent or that may be difficult to discern using traditional methods. Weak signal analysis methods can include techniques such as trend analysis, anomaly detection, and horizon scanning. The goal of using weak signal analysis methods in emerging technology detection is to provide early insights into new technologies that can aid in the identification and development of these technologies before they become mainstream.

Weak signal analysis can be considered another type of emerging technology detection method that has been used recently. The recent definition of the weak signal based on the previous definitions is proposed by van Veen and Ortt (2021): “A perception of strategic phenomena detected in the environment or created during interpretation that are distant to the perceiver’s frame of reference.” This concept can be used in emergence

detection because these weak signals may be potential signals of future trends and technologies.

Yoon (2012) first introduced the concept of weak signal analysis to detect business opportunities, trying to identify weak signal topics using text mining based on keywords. Its data type was Web news articles, and its domain of research case study was solar cells. Using term and document frequencies, Yoon (2012) described a weak signal as a term with low term and document frequencies and a high growth rate. This study also described a strong signal as a term with high term and document frequencies and a high growth rate.

Griol-Barres et al. (2020) combined three data types from scientific, journalistic, and social sources to detect weak signals in the field of remote sensing. It uses three data sources; papers from Science Direct, newspapers from New York Times, and social media from Twitter to capture future changes using weak signals in the targeted field.

Ebadi et al. (2022) applied recent concepts of weak signal analysis to the early detection of technology emergence. They mixed a deep learning and NLP (Natural Language Processing) approach to detecting keywords with a weak signal analysis to identify emerging terms as future signs early in the field of "hypersonic" using scientific publications.

H. Xu et al. (2021) attached the concept of weak signal in emergence detection to the uncertainty and ambiguity attribute and tried to measure it into that attribute in the process of emergence detection. They suggested that capturing weak signals can help the process of early identification of emerging technologies.

Weak signals are considered to be emerging trends that are yet to catch the eye of experts and growing areas that show high dynamics but are not widespread yet (Nazarenko et al. 2022). Therefore, weak signal analysis can play an important role in the process of emerging technologies detection. It can potentially even become a new trend in the identification of emerging technologies as some more advanced techniques such as quantum computing are being used in detecting future weak signals (Griol-Barres et al. 2021).

Broadly, weak signal analysis can be useful for identifying emerging technologies that are still in the early stages of development, but it should be used in conjunction with other methods and approaches to get a more complete understanding of the field and its

trends.

2.2.6 Hybrid Approaches

Hybrid methods in detecting emerging technologies refer to the combination of multiple techniques, methods, and algorithms from different fields, such as machine learning, lexical, bibliometrics, and indicator methods, to identify and track new or developing technologies (Rotolo et al. 2015). The goal of using hybrid methods is to create a more comprehensive and accurate approach by leveraging the strengths of different methods and to overcome the limitations of a single method. By combining multiple methods, hybrid methods can provide a more holistic view of the technology landscape, allowing for a more accurate identification of emerging technologies and the trends and patterns associated with them. Hybrid methods can also be used to increase the efficiency of the analysis process by allowing the use of multiple data sources, and to increase the robustness of the results by combining different types of information.

Ávila-Robinson and Miyazaki (2013) used bibliometric indicators to detect technological emergence. Using the Thomson Reuters/ISI Science Citation Index Expanded database and their citations and references, they integrated bibliometric, social network analysis and multivariate statistical methods. Q. Wang (2018) used bibliometrics and indicators of growth of the number of publications, the novelty of the topic, coherence of the cluster of a topic, and the number of citations as the scientific impact to detect an emerging research topic.

J. Garner et al. (2017) used a set of indicators to evaluate the emergence of the terms in terms of novelty, growth, community, and persistence, which was a combination of lexical and indicator methods; their method was called Emergence Score (EScore). Carley et al. (2017) used the same EScore method but evaluated the effects of scale and domain on the persistence of an emerging topic. Carley et al. (2018) elaborated on the EScore method more thoroughly, implementing it on a dye-sensitized solar cells (DSSCs) dataset, and found the emerging terms, authors, and affiliations in this field. Porter et al. (2019) implemented the EScore method and revised it to identify the emerging terms and key players, as well as high-priority research papers and patents.

Ranaei et al. (2020) compared the EScore method with other approaches, concluding that EScore provided a robust, holistic view of technological emergence by integrating term frequency, community size, and origin parameters. However, the method had limitations in capturing highly niche or less-researched areas. They recommended using EScore in combination with other methods, like term counting or LDA, to get a more comprehensive understanding of emerging technologies.

Overall, hybrid methods can be a powerful tool for detecting emerging technologies, but it is important to recognize their complexity and the need for expert knowledge to implement and interpret the results.

2.2.7 Analysis of the Literature and Research Gaps in Approaches to Emergence Detection

A comprehensive list of emergence detection methods is presented in Table 1. The categorization of these methods proves challenging and intricate due to the inherent difficulty of assigning a singular category to each method without encountering overlap with other categories. Nonetheless, efforts have been made to categorize each paper's method, drawing upon the previously established definitions and descriptions provided in preceding sections.

A list of different emergence detection methods with their pros and cons (as mentioned by the authors) can be seen in Table 1.

Table 1. List of different emergence detection methods

Papers	Emergence Detection Method	Pros	Cons
Mejia & Kajikawa (2020)	Bibliometric	Clear identification of influential works	Limited to citation-based networks
Kim & Bae (2017)	Bibliometric	Effective for citation-based trend detection	Limited to research-focused trends

Iwami et al. (2014)	Bibliometric	Clear methodology for mapping trends	Limited by citation network bias
Fujita et al. (2014)	Bibliometric	Effective for detecting well-cited research	May overlook emerging but less-cited fields
Yoon & Kim (2012)	Bibliometric	Effective in technology classification	May miss emerging niche technologies
Kajikawa et al. (2008)	Bibliometric	Effective for identifying influential research	Misses emerging but less-cited technologies
Shibata et al. (2008)	Bibliometric	Simple and effective	Limited to citation data
Huang et al. (2021)	Hybrid	Combines strengths of multiple approaches	Complexity in integration
X. Liu & Porter (2020)	Hybrid	Broader scope through hybrid approaches	Difficult to implement due to complexity
Ranaei et al. (2020)	Hybrid	Comprehensive view of technological emergence	Less effective for niche technologies
Li et al. (2019)	Hybrid	Comprehensive in scope	Complexity in combining different methods
Porter et al. (2019)	Hybrid	Combines multiple data sources	High rates of noise in final results
Q. Wang (2018)	Hybrid	Multiple sources increase accuracy	Resource-intensive

Carley et al. (2018)	Hybrid	Easy to understand and apply	High rates of noise in final results
H. Xu et al. (2021)	Indicator	Focused on specific technology indicators	May miss broader, context-based signals
Guderian (2019)	Indicator	Simple and transparent methodology	Limited to selected indicators
Z. Wang et al. (2019)	Indicator	Effective in specific technology areas	Limited to patent-driven detection
Moehrle & Caferoglu (2019)	Indicator	Focuses on specific indicators	May miss emerging trends outside data sources
Guo et al. (2011)	Indicator	Simple methodology	May miss broader trends
Shibata et al. (2011)	Indicator	Clear methodology	Limited by citation network bias
Weismayer & Pezenka (2017)	Lexical	Easy to implement	Limited to text analysis
Joung & Kim (2017)	Lexical	Efficient for analyzing text corpora	Limited to keyword matching
Furukawa et al. (2015)	Lexical	Easy to interpret	Limited by the scope of lexical data
Zhou et al. (2021)	Machine Learning	Scalable with large datasets	Requires high computational power

S. Xu et al. (2021)	Machine Learning	Adapts well to large datasets	Dependent on data quality
Jang et al. (2021)	Machine Learning	Handles unstructured data well	Requires extensive training data
Choi et al. (2021)	Machine Learning	Handles large text corpora well	Can overlook subtle emerging signals
Ma et al. (2021)	Machine Learning	Effective for large datasets	Resource-intensive implementation
Zhou et al. (2020)	Machine Learning	Robust performance in structured data	Limited to specific data formats
Altuntas et al. (2020)	Machine Learning	Efficient at processing large volumes of data	Data-dependent performance
Kwon & Geum (2020)	Machine Learning	High scalability	Requires high-quality data for best results
S. Xu et al. (2019)	Machine Learning	Handles large text datasets well	Requires extensive training data
Zhou et al. (2019)	Machine Learning	Effective in large dataset applications	Requires high computational resources
Ebadi et al. (2022)	Weak Signal	Effective in early-stage detection	Returns some noisy terms

2.3 Data Sources Used to Detect Emergence

There are different data sources used for the identification of emerging technologies and topics. From science and technology (S&T) sources including papers and patents as the

typical sources of emergence detection in the literature to novel data sources such as altmetrics, news, or funding. Various online platforms are used to measure the dissemination of research results using altmetrics (Akella et al. 2021). Based on information in Table 2, from our 34 main research in emergence detection studies, 18 research works used papers as their main source of emergence, 14 used patents, one used both, and one used online posts. As seen in Table 2, one research gap in the literature on emergence detection could be using other types of data such as altmetrics, news, social media, or funding as the main source of data emergence.

2.3.1 Research Papers

Research papers are considered as the main source of the data in the process of emergence detection in the literature. This can include articles and journal papers or conference papers. One of the advantages of using papers to detect emerging technologies is that they are perceived to be upstream of patents (Liu and Porter 2020); therefore, they can identify the emerging topics sooner.

Some of the works that use papersas their main data source of emergence try to detect emerging research topics on the basis that using papers as the source might lead to the finding of emerging research topics instead of emerging technologies. You can see works that used research papers as their main source of emergence detection in Table 2.

2.3.2 Patents

Many researchers used patents in the process of emergence detection, especially detecting emerging technologies. As patents are more closely related to technology than research papers, some might prefer evaluating patents as their main data source of emergence. You can see works that used the patents as their main source of emergence detection in Table 2.

2.3.3 Supplementary Data

In this thesis, the concept of supplementary data in the process of emergence detection will be discussed. Supplementary data encompasses information from which emerging topics or

technologies are not directly extracted. Rather, this data serves as a supplementary resource in the process of detecting or assessing emergence. You can see the supplementary data in Table 2.

Four papers in the literature worked on supplementary data. Furukawa et al. (2015) detected potential emerging topics from papers but also used names of conference sessions to detect and assess emerging topics. Li et al. (2019) used tweets as a supplementary data source to patents for detecting emerging technologies and development trends. H. Xu et al. (2021) used patents as supplementary data to papers to better evaluate the potential for the prominent impact of the candidate technologies. Zhou et al. (2021) used web articles to evaluate the social and technological impact of the potential technologies extracted from patents.

2.3.4 Domain of Datasets

There are many different science and technology domains studied in the literature on emergence detection by researchers, including dye-sensitized solar cells (DSSC), which is the most frequently researched topic compared to other domains, such as stem cells, 3D printing, and even hypersonic technology.

Table 2. Research gap of different data types of emergence detection

Papers	**Data Emergence Source**	**Supplementary Data**	**Domain of Datasets**
(Ebadi et al. 2022)	Paper	×	hypersonic
(Zhou et al. 2021)	Patent	Web Article	CNC machine tool
(S. Xu et al. 2021)	Paper	×	synthetic biology
(H. Xu et al. 2021)	Paper	Patent	stem cells
(Jang et al. 2021)	Online posts	×	food processing
(Huang et al. 2021)	Paper	×	Information Science

(Choi et al. 2021)	Patent	×	Batteries for electric vehicles
(Ma et al. 2021)	Patent	×	Dye sensitized solar cell
(Zhou et al. 2020)	Patent	×	×
(X. Liu and Porter 2020)	Paper	×	NEDD (Nano Enabled Drug Delivery), Non-Linear Programming, DSSCs
(Altuntas et al. 2020)	Patent	×	dental implant technology
(Kwon and Geum 2020)	Patent	×	electric digital data processing
(Ranaei et al. 2020)	Paper	×	LEDs and flash memory
(Mejia and Kajikawa 2020)	Paper and Patent	×	energy storage
(S. Xu et al. 2019)	Paper	×	gene editing
(Zhou et al. 2019)	Paper	×	3D printing
(Guderian 2019)	Patent	×	smart houses
(Z. Wang et al. 2019)	Patent	×	3D printing
(Moehrle and Caferoglu 2019)	Patent	×	camera technology
(Li et al. 2019)	Patent	Tweets	perovskite solar cell
(Porter et al. 2019)	Paper	×	NEDD, Non-Linear

			Programming, DSSCs, Big Data
(Q. Wang 2018)	Patent	×	×
(Carley et al. 2018)	Paper	×	DSSCs and Nonlinear Programming
(Weismayer and Pezenka 2017)	Paper	×	marketing and tourism
(Joung and Kim 2017)	Patent	×	electron transfer in glucose biosensors
(Kim and Bae 2017)	Patent	×	wellness care
(Furukawa et al. 2015)	Paper	Conference Session	Web-based technology
(Iwami et al. 2014)	Paper	×	Fluorescent protein, Cryptology, Quasicrystal, iPS cell
(Fujita et al. 2014)	Paper	×	gallium nitride, complex networks, and nano-carbon
(Yoon and Kim 2012)	Patent	×	organic photovoltaic cells
(Guo et al. 2011)	Paper	×	RNAi, Nano, h-Index, and Impact Factor
(Shibata et al. 2011)	Paper	×	regenerative medicine
(Kajikawa et al. 2008)	Paper	×	energy

(Shibata et al. 2008)	Paper	×	gallium nitride (GaN) and complex networks

2.3.5 Evaluation of Data

Scientific papers are a reliable source of information about new and emerging technologies, as they have undergone peer review and have been deemed credible by experts in the field (Rotolo et al. 2015). They provide detailed information about the technology, including its development, potential applications, and limitations, which can be useful for an automated process. Many scientific papers are now available in digital format, making it easier to collect and process large amounts of data from this source.

However, not all emerging technologies are necessarily published in scientific papers, so relying solely on this source may lead to missing some important developments. Papers can be technical and difficult to understand for those without a background in the field, which can make it more difficult to automatically extract useful information (Park and Yoon 2018).

Patents are a good source of information about new technologies that are being developed and commercialized. Patents provide information about the inventors and companies involved in the technology, which can be useful for understanding its potential impact and commercial potential. Patents are often filed before the technology is commercially available, making them a good source for identifying emerging technologies.

However, patents are often written in legal language and can be difficult to understand for those without a background in patent law (Park and Yoon 2018). Patents may furthermore not provide as much detail about the technology as scientific papers.

It is important to note that both patents and scientific papers are different sources of data and have their own advantages and disadvantages. Patents capture applied, market-driven innovations, while scientific papers provide insights into theoretical advancements and research trends. By combining them, we can identify both the development of new technologies and the underlying research, offering a fuller picture of technological emergence (H. Xu et al. 2021).

It can be valuable to use supplementary data such as funding information, altmetrics,

online posts, or newspapers in the process of emerging technology detection, as they may offer additional perspectives not captured by scientific papers or patents alone. Although not widely adopted in the literature, these sources could enhance the understanding of technology trends. Some insights provided by these aspects of supplementary data follow:

- Funding data can provide information about which technologies are receiving investment, which can be a good indicator of their potential impact and commercial potential.
- Altmetrics, such as social media mentions, can provide insight into which technologies are receiving attention and interest from the public and stakeholders, which can be an indicator of their potential impact.
- Online posts and news articles can provide information about the real-world applications and uses of technologies, as well as any challenges or barriers to adoption.
- Combining multiple data sources can give a more comprehensive understanding of the technology's potential impact, commercial potential and its current state.
- There are still many domains that have remained untouched. One of them is artificial intelligence (AI) which can be a suitable candidate for future work.

Furthermore, by using different sources of data, it can help to account for the limitations of any one data source. For example, patents may not provide as much detail about the technology as scientific papers or may be filed before the technology is commercially available, whereas funding and altmetrics data can provide a more real-time picture of the technology's current state and its potential impact.

2.4 **Burst Detection**

In this section, we briefly summarize and look at the literature of burst detection. We mainly discuss the entire concept of burst detection as well as the details of the methods related to burst detection found in the literature.

In the literature, "bursty terms" refer to terms that experience a sudden increase in usage and popularity within a certain time frame, contrasting with those that are consistently popular (Katsurai & Ono, 2019). Early methods of burst detection involved segmenting the

corpus into topics and tracing changes in their popularity across different years. One of the foundational methods is Kleinberg's burst detection algorithm, which identifies bursts of activity in streams of data based on the frequency of certain terms over time (Kleinberg, 2002). This algorithm has been used in various contexts, including social media platforms like Twitter and news feeds (Diao et al., 2012; Fung et al., 2005).

The idea of burstiness has been explored in various fields, including finance (Murphy 1999), disaster or biological studies (Ruzzo and Tompa 1999), traffic management (A. Zhou et al. 2005), and information technology (Mane and Börner 2004). Different applications typically use specific burst models and detection techniques. A large body of research has emerged on burst detection, much of which is influenced by the work of Kleinberg (Kleinberg 2002) and Shasha (Zhu and Shasha 2003). Kleinberg's approach models bursty streams using an infinite-state automaton, where each state represents a data emission rate governed by an exponential distribution of time gaps between data arrivals. Shasha's method defines bursts based on fixed-length sliding windows to enable efficient monitoring of elastic bursts in higher-dimensional data. Subsequent research has proposed improved definitions and detection methods, addressing general-purpose use, multi-sequence data, and computational efficiency. For instance, Zhou introduced the concepts of increasing and decreasing bursts, noting that the ratio of aggregate values in consecutive subsequences of the same length can help identify bursts (A. Zhou et al. 2005).

T. Chen et al. (2006) incorporates a lasting factor and an abrupt factor into the standard definition of a burst, aiming to quantify the aggregate value within a time window and the growing discrepancy between two points within that window in practical scenarios. Lappas et al. (2009) details a new search technique that detects term burstiness in document sequences by applying principles from discrepancy theory. This paper suggests a parameter-free, linear-time method for pinpointing the time intervals where a given term exhibits its highest burst.

A commonly used technique for extracting topics from a collection of documents is Latent Dirichlet Allocation (LDA), introduced by Blei et al. (2003). LDA has become integral to understanding thematic structures in large text datasets. However, despite its widespread use, LDA has some drawbacks, such as its lack of interpretability and difficulty in linking topics across different time periods (Tattershall et al., 2020). This limitation makes

it less ideal for detecting bursts in evolving research areas.

Kleinberg's burst detection algorithm, though effective in real-time environments such as social media, is less applicable to scientific literature because scientific papers are not continuously published but rather in batches (Tamura and Kitakami 2014). Furthermore, changes in scientific research happen over longer time scales compared to the real-time nature of tweets or news updates. To address this limitation, alternative burst detection methods have been proposed. For example, He and Parker (2010) adapted stock market analysis models to detect bursts in scientific documents, allowing for more effective analysis of term emergence in slower, cumulative datasets.

Thus, burst detection can be a tool in understanding the dynamics of emerging technologies, with adaptations that span various data types and domains. While challenges such as linking bursts across time periods and managing real-time versus slower streams remain, continued improvements in methods ensure more robust applications in detecting and forecasting emerging trends.

2.5 Discussion on Literature Review

The literature review has extensively covered various methodologies and data sources prevalent in the field of emerging technology detection along with the burst detection method that can be useful in the process of emerging technology detection. Despite the comprehensive review of existing studies, it was observed that the literature on emerging technology detection has increasingly underscored the significance of automating the detection process to minimize subjectivity and enhance reproducibility. Traditional methods predominantly rely on manual interventions and expert judgments, introducing a risk of subjectivity that may skew the detection and evaluation of emerging technologies (Rotolo et al., 2015). Moreover, these methods often lack scalability and flexibility to adapt to rapidly changing technology landscapes, which is critical given the exponential growth of data and technology domains.

Further compounding the issue, traditional approaches typically do not employ quantifiable metrics that facilitate an objective assessment of their performance. This absence hinders the ability to gauge the effectiveness of the technology detection process

systematically, thereby limiting the predictive validity of these methods concerning future technology trends (Porter et al., 2002). Such gaps underscore the necessity for innovative methods that incorporate robust, quantifiable metrics to enhance predictive accuracy and reliability.

In response to these challenges, this thesis proposes an innovative method leveraging burst detection and machine learning techniques to identify and predict the trends of emerging technology terms within the AI sector. Unlike conventional methods, the proposed approach utilizes burst detection algorithms to automate the initial detection of emerging terms, significantly reducing the need for manual curation and thereby mitigating the risk of human biases. This automation is pivotal in enhancing the objectivity and efficiency of the detection process.

Moreover, by integrating machine learning models with burst detection, the proposed method not only identifies emerging terms but also predicts their future trajectory based on historical data patterns. This integration facilitates the generation of quantifiable performance metrics, such as accuracy and recall rates, which are critical in evaluating the effectiveness of the detection method. Such metrics provide a solid foundation for continuously improving the detection process through systematic feedback and adjustment.

In summary, the proposed method addresses the identified gaps in traditional emerging technology detection approaches by enhancing objectivity, scalability, and predictability through the innovative use of burst detection and machine learning techniques. This approach not only contributes to the theoretical advancements in technology detection methodologies but also offers practical applications for stakeholders aiming to harness the potential of emerging technologies effectively.

3 Data

Two types of data are used in this research; we used different sources and processes to collect and prepare the data depending on the type thereof. In the following paragraphs, each data type and its process are discussed.

3.1 Patents

Patent data were sourced from the Lens database. We conducted a query search using the terms "artificial intelligence," "deep learning," and "machine learning" within titles and abstracts, spanning from 1990 to 2023. For each patent, we extracted both the abstract and title, and yearly vectorization was performed. Additionally, we employed two distinct thesauri to eliminate English stopwords and common words. Unrelated terms outside the AI domain were also filtered out. This process resulted in a dataset of 1.2 million AI-related patents from 1990 to 2023.

3.2 Papers

We also compiled academic papers base on the previous databases in the literature to better compare the results of this method to the previous ones. We used the comprehensive dataset from DBLP, a well-regarded computer science bibliography hosted by Trier University, Germany. After gathering the data and filtering non-English abstracts, the final version of this dataset includes 2.6 million articles from 1988 to 2017.

4 Methodology

In this chapter, we will provide a detailed explanation of the techniques and approaches used, including burst detection, machine learning, and deep learning methods, to predict the sustainability of emerging technology terms. Moreover, we analyze the metrics of machine learning and deep learning approaches in order to assess the exent of these approaches's ability to predict the future sustainability of the emerging technology terms. In the context of emerging technologies, we refer to sustainability as the ability of emerging technologies to continue developing and being adopted over time, rather than fading out. You can see an overview of the conceptual flow of the proposed methodology in Figure 2.

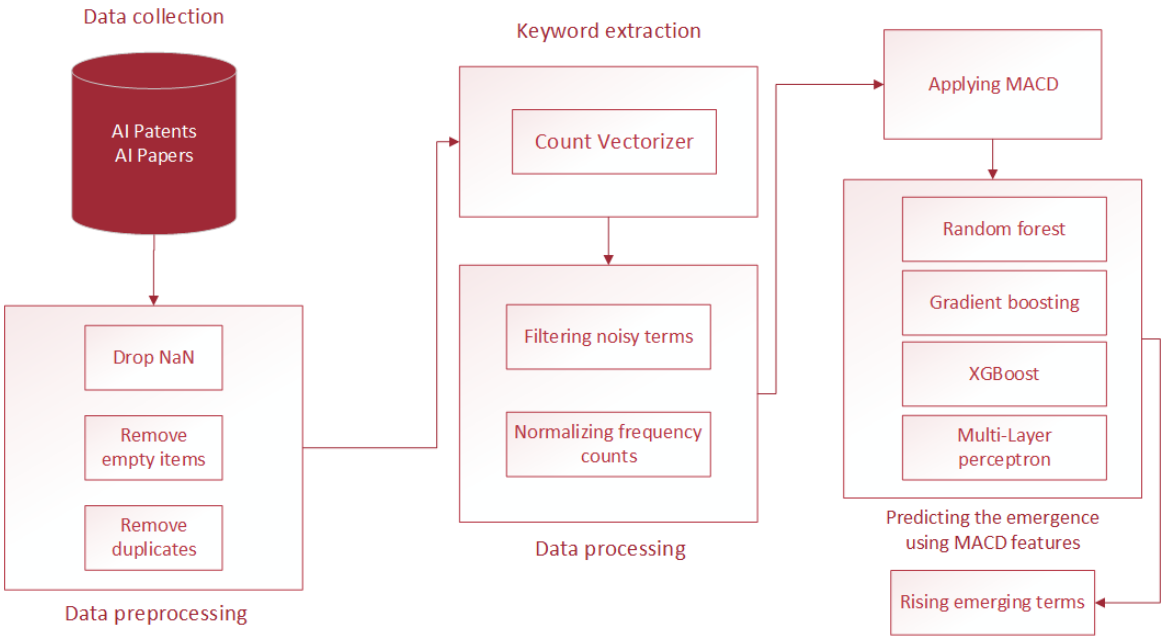

Figure 2. An overview of the conceptual flow of the proposed methodology for detecting emerging technologies using machine learning and burst detection

4.1 Keyword Extraction Using the Count Vectorizer

In the first step, we will identify the key phrases using a keyword extraction method (Ghaemmaghami et al. 2022). The process begins with the use of the count vectorizer and stop-words to extract key phrases used every studied year. We also tried to exclude general

terms in two steps: firstly, the terms that are general English terms have been removed using two English word thesauri; secondly, any unmeaningful term in the context of studies concerning AI have also been removed using a list of terms related to this field. The remaining terms used each year have been added to create a yearly vector of terms.

4.2 Filtering and Normalization

We faced two main challenges at this step of our work, after the keyword extraction. First, the output terms could be noisy and include many terms that appear in very few records. Second, inasmuch as the AI domain is undergoing an increase in popularity and interest, new terms are becoming more prevalent, not because they are more emerging or bursty but because of the increase in the data size. To overcome these challenges, we followed the approach proposed in Tattershall et al. (2020), consisting of two steps: 1) remove any terms that have been absent from more than 0.02% of the body text or abstract for at least 3 successive years, and 2) normalize the frequency counts for each document twice, first by dividing their total number by each year's number of documents, then by the total number of tokens per document. The normalized frequency count, called prevalence, was used as the main input in the calculations in the following sections.

4.3 Moving Average Convergence Divergence

The Moving Average Convergence Divergence (MACD) is a technical analysis tool that uses exponential moving averages (EMAs) to smooth out stock price fluctuations and reveal underlying trends (He and Parker 2010). They are a type of moving average that gives more weight to the most recent data points in a time series. In this work, we use the MACD notions described by He and Parker (2010). For a time span *n*, and a time series variable such as price of a stock (or, in our case, the frequency of a term in bibliometric record) in time *t* as $\mathrm{y}(t_i)$:

$$\mathrm{EMA}(t_i) = \mathrm{EMA}(t_{i-1}) + (2/(\mathrm{n} + 1)) * (\mathrm{y}(t_i) - \mathrm{EMA}(t_{i-1})) \quad (1)$$

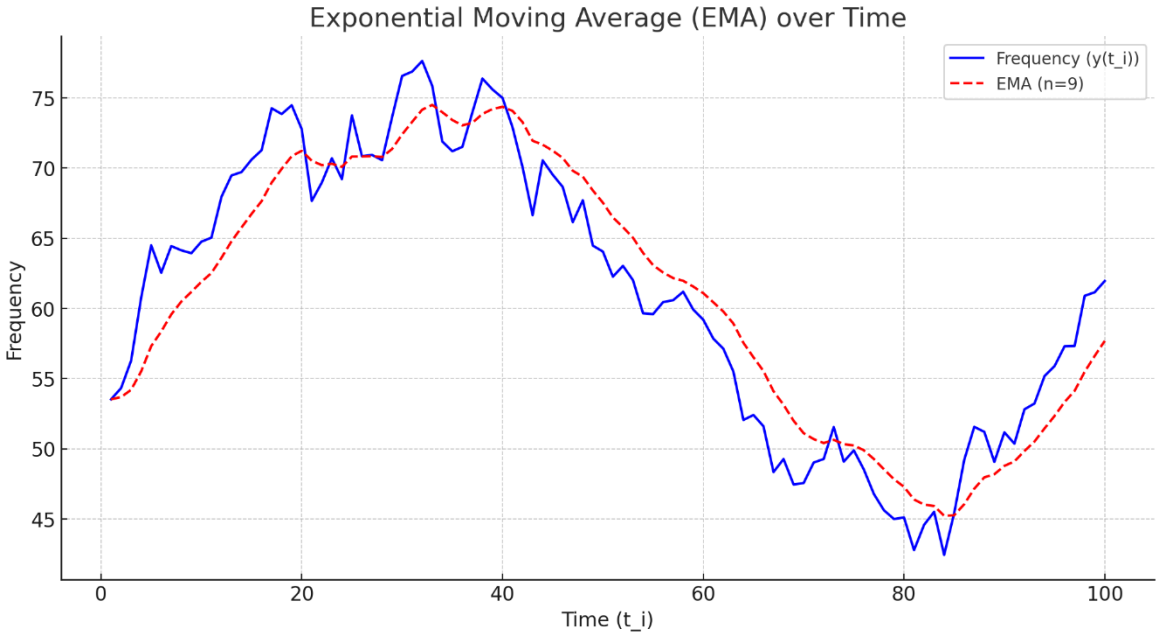

Figure 3. Exponential Moving Average (EMA) Over Time: Demonstrating the EMA Calculation for Stock Prices

Figure 3 illustrates the concept of EMA by plotting its calculation against a time series, highlighting how it smooths out stock price fluctuations and reveals underlying trends.

The MACD is calculated by subtracting a long EMA from a short EMA, resulting in the MACD line (Eq. 2). Long EMA covers more time spans in comparison to short EMA.

$$\mathrm{MACD}(n_1, n_2) = \mathrm{EMA}(n_1) - \mathrm{EMA}(n_2) \quad (2)$$

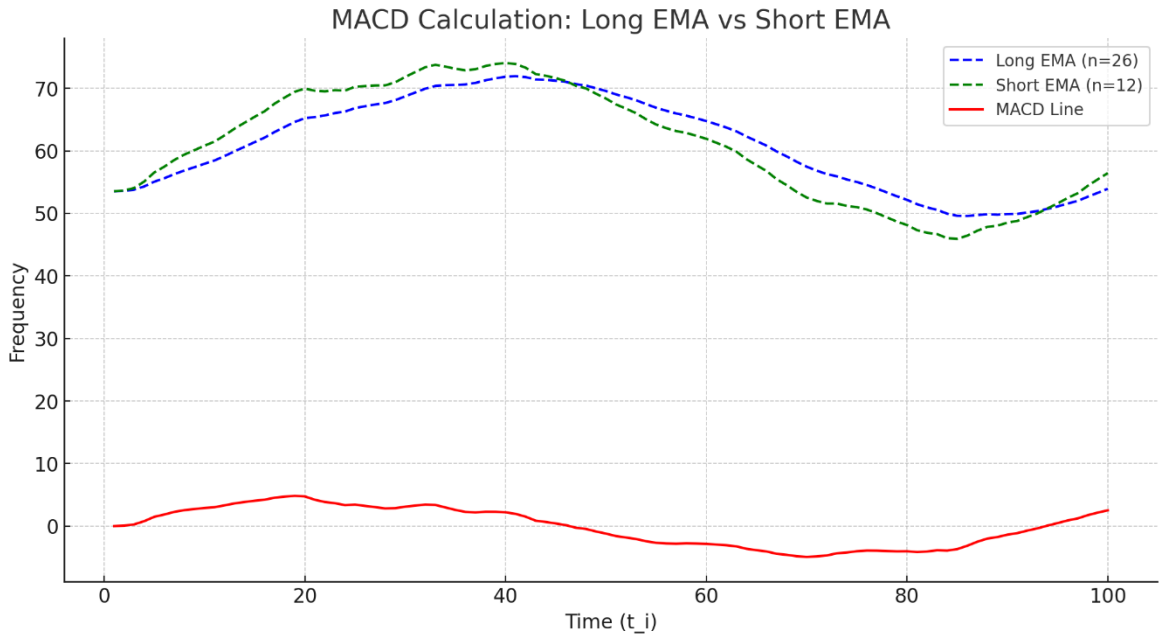

Figure 4. MACD calculation: Long vs short exponential moving averages: Visualizing the components of the MACD line

Figure 4 focuses on the MACD line's formulation, showcasing the dynamic interaction between the long and short EMAs to demonstrate how their difference forms the basis of MACD analysis.

This MACD line is then averaged with an EMA of a third span, creating the signal line (Eq. 3). The signal line is the smoothed MACD line and helps to identify buy and sell signals or the points that the trend is changing to upward or downward by reducing the noise and making it easier to see the trend.

$$\text{signal}(n_1, n_2, n_3) = \text{EMA}(n_3)[\text{MACD}(n_1, n_2)] \quad (3)$$

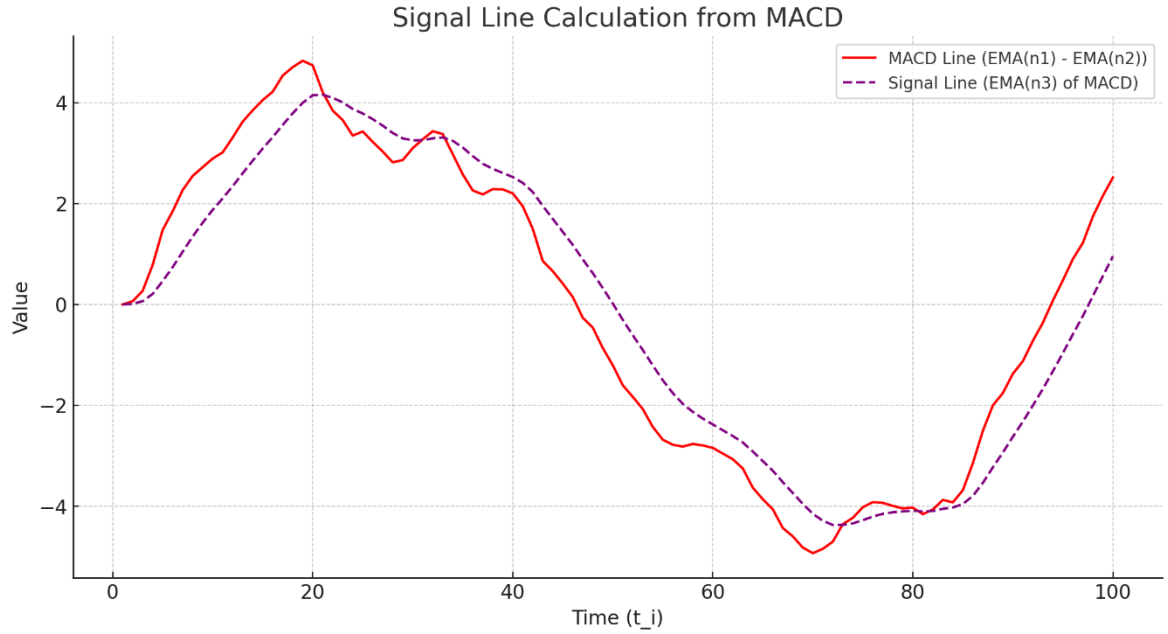

Figure 5. Signal line derivation from MACD: Showcasing the smoothing of the MACD line

In Figure 5, we delve into the smoothing of the MACD line, illustrating how the Signal Line is calculated by applying an EMA to the MACD line, and emphasizing its role in highlighting potential buy and sell signals.

In Figure 6, a histogram is used as an indicator of price acceleration by comparing the MACD line with the signal line (Eq. 4). When there is a positive trend, the histogram is positive. Therefore, a change in the value represented by the histogramfrom negative to positive can be sign of changing a trend from negative to positive.

$$\text{histogram}(n_1, n_2, n_3) = \text{MACD}(n_1, n_2) - \text{signal}(n_1, n_2, n_3) \quad (4)$$

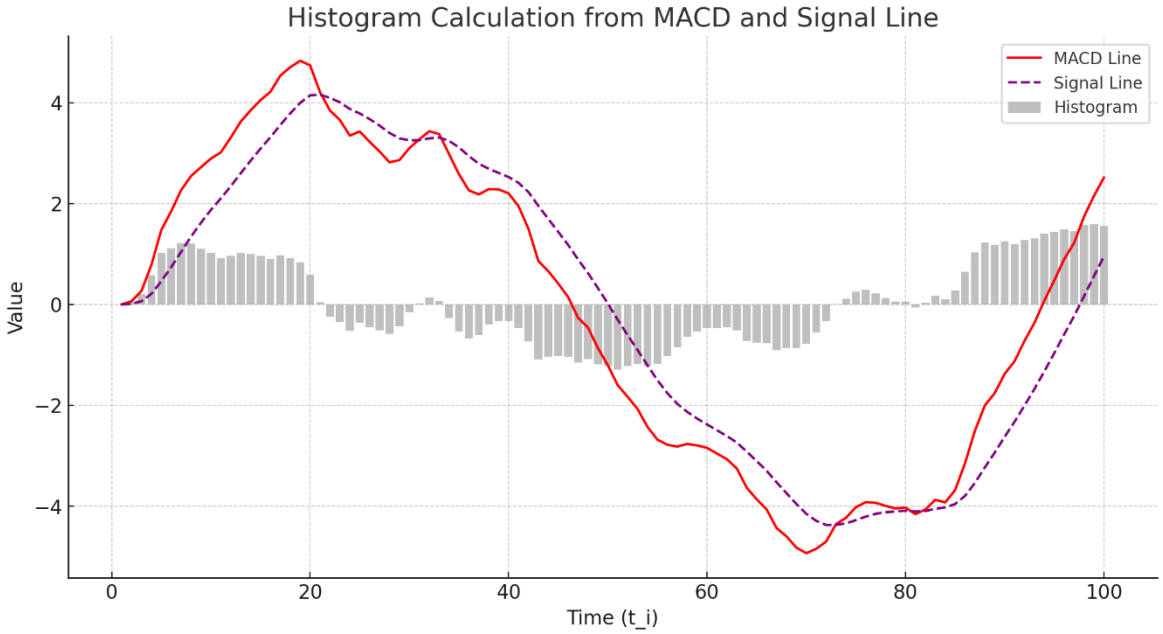

Figure 6. Histogram analysis in MACD: Illustrating the difference between MACD and signal line as an indicator of trend changes

The Figure 6 aims to demonstrate the significance of the MACD Histogram by comparing it with the MACD and Signal lines, underscoring how changes in the histogram can indicate shifts in market trends.

The MACD has been applied to analyze scientific data, such as the frequency of Medical Subject Headings (MeSH) in scientific papers instead of the stock price over time (He and Parker 2010). A modified version of it (presented in Figure 7) is also used to identify bursty terms in the computer science field (Tattershall et al. 2020).

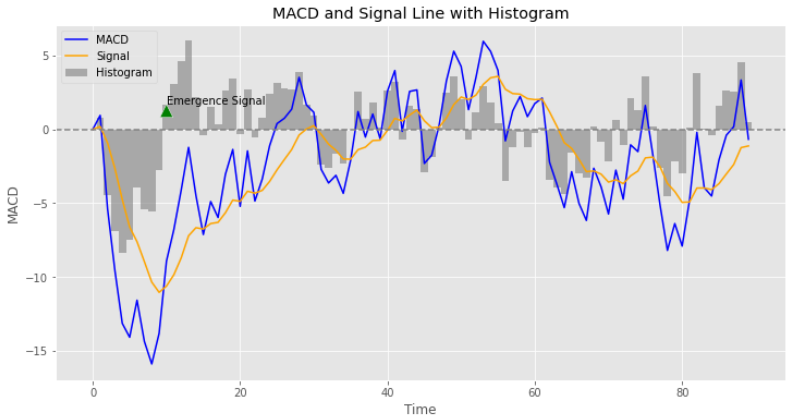

Figure 7. MACD, Signal Line, and Histogram and how they can signal an increase in the time series variables. In our case, this is a sample random data to show the "emergence signal" based on the frequency of a term in scientific papers.

In Figure 7, we tried to visually show the main indicators of MACD, Signal Line, and Histogram in the context of time series in scientific papers.

4.4 Applying MACD

We selected a range of (6, 12, 3) as our parameters for the moving average spans (n_1, n_2, n_3). Different methods use different burstiness calculations. The raw value of the histogram was used by He and Parker (2010) as a measure of burstiness, while Tattershall et al. (2020) used the square root of the historical maximum prevalence as the scaled factor to calculate burstiness. We followed Tattershall et al. (2020) as it produces more consistent results. The burstiness is calculated based on the prevalence of a specific term *w* in time *t*, p(w, t), as follows:

$$Burstiness[n_1, n_2, n_3](p(w,t)) = histogram[n_1, n_2, n_3](p(w,t)) / \sqrt{max(p(w,t))} \quad (5)$$

We also considered the burstiness investigation period to be a changeable parameter, n_4. It refers to the number of previous time spans that should be considered in the process of taking MACD features for the bursty terms. In the literature, it has been considered unchangeable and assigned 8 for n_4. Through experimentation, we found that 8 and 9 are the best n_4 values for papers and patents, respectively, in our datasets.

4.5 Predicting the Emergence Using MACD Features

The next step is to build a supervised learning model that receives the term and its features as the input and produces a label as the output. Following Tattershall et al. (2020), we built the model based on MACD features. However, the data splitting method we use is a time-based split, which differs from the approaches used in previous studies. Previous studies employ random sampling methods for splitting the dataset into training and test sets. For instance, random cross-validation techniques, where data points are randomly selected to form the training and testing sets, are prevalent in studies like in Tattershall et al. (2020). These methods, while effective for general statistical validation, might not account adequately for the temporal nature of data in technology forecasting where the order of data points (i.e., chronological order) plays a crucial role.

According to these points, we formulate our methodology by the following steps. The prediction interval *I* indicates the prediction time window, i.e., the number of years ahead the prediction is made (e.g., 3). The algorithm works as follows, for each of the following

years, y_i:

1. Consider the range of data from [(i+n_2+n_4-1), N].
2. Consider the whole set of data D($y_{(i+n_2+n_4-1)}$, y_N). (N is time span in our data)
3. Apply burst detection to D($y_{(i+n_2+n_4-1)}$, y_N) and select all terms with burstiness levels above a certain level that we choose as a parameter.
4. Calculate MACD, histogram, standard deviation, minimum and maximum values as the X component (the input).
5. Calculate if the smoothed value of term prevalence during y_{i+I} is higher or lower than the prevalence during y_i, as the Y component (the label).
6. Assign X_train and Y_train by time split to X($y_{(i+n_2+n_4-1)}$, y_N-1) and Y($y_{(i+n_2+n_4-1)}$, y_N-1)
7. Assign X_test and Y_test by time split to X(y_N) and Y(y_N) (latest available year)

4.6 Baseline Machine Learning Models

Prior works (Balili et al. 2017; He and Parker 2010; Tattershall et al. 2020) used a tree-based method for predicting the popularity of clusters or terms. We trained and built a Random Forest classifier as the baseline. We tuned the hyperparameters of the Random Forest model for each data type and applied the best maximum depth and number of estimators to achieve the highest possible accuracy.

While previous studies primarily utilized Random Forest classifiers to predict the popularity of terms, this research extends the comparative framework by incorporating Gradient Boosting and XGBoost models to compare the performance of these conventional machine learning methods and identify which models to prioritize for further use. These additional models were meticulously tuned for hyperparameters such as maximum depth and number of estimators to enhance accuracy, thus providing a comprehensive evaluation of their effectiveness in identifying emerging trends.

4.7 Construction of a Neural Network Classifier

We built a Multi-Layer Perceptron (MLP) neural network classifier and compared the results with the baseline Random Forest classifier. The use of the MLP classifier based on MACD features will enable the capturing of complex time series problems, such as stock market prediction and emerging technologies detection; deep learning models may work more effectively than conventional machine learning methods, given their ability to capture more complex relationships. It is important to optimize the MLP's architecture and determine the best configuration, because the structure of a neural network (i.e., the number of layers and units per layer) can significantly affect its performance. Therefore, we will find the optimal number of layers and nodes by running a series of experiments to maximize performance.

The introduction of a MLP neural network classifier represents a further departure and advancement over the conventional methods cited in earlier studies. Unlike traditional tree-based methods that might struggle with the complexity and non-linearity of time-series data often found in technology forecasting, the MLP classifier is specifically adapted to handle these complexities. This adaptability is crucial for effectively modeling and predicting dynamics such as those seen in emerging technologies and stock market fluctuations. To ensure the neural network's optimal performance, extensive experimentation was conducted to determine the most effective network architecture, including the appropriate number of layers and nodes. This approach not only aligns with the methodologies from foundational literature but significantly augments them by leveraging deeper insights into data patterns, thereby offering improved accuracy, efficiency, and applicability in forecasting emerging technologies. These enhancements highlight the thesis's contributions to advancing the field of technology detection and prediction, underscoring the benefits of integrating sophisticated machine learning techniques to better capture and analyze the nuanced behaviors of technological evolution.

4.8 Model Evaluation

In the context of a binary classification problem, where outcomes are categorized as either positive (p) or negative (n), four potential outcomes need consideration:

- True Positive (TP)

- False Positive (FP)
- True Negative (TN)
- False Negative (FN)

If the prediction outcome is p and the actual value is also p, it is classified as TP. Conversely, if the true value is n, it is classified as FP. When both the prediction and true values are n, it is categorized as TN. FN occurs when the true value is p, but the prediction is n. Various metrics can be computed based on TP, FP, TN, and FN, providing valuable insights into the evaluation of machine learning classifiers.

4.8.1 Precision

Precision, also known as positive predictive value, is defined as the ratio of true positive predictions to all positive predictions, as illustrated in the following equation. It serves as a relevant evaluation metric, especially in scenarios where the cost of FP is high, such as in email spam detection models.

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

4.8.2 Recall

Recall gauges how effectively a machine learning model predicts instances of the actual positive class. The following equation depicts recall as the proportion of TP predictions out of all actual positive examples. This metric is particularly useful when the cost associated with FN is substantial, as seen in applications like cancer detection models.

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

In our problem, it is more important not to miss rising terms than identifying false rising terms. Therefore, out of the two metrics, recall is the more important one.

4.8.3 F1 Score

The F1 score, also referred to as the F-score or the F-measure, quantifies the weighted

average of precision and recall. The following equation represents the F1 score as the harmonic mean of precision and recall, providing a balanced assessment. The F1 score ranges from 0 to 1.

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (8)$$

4.8.4 F1-score Weighted Value

The F1-score is especially valuable when dealing with imbalanced classes. The F1-score weighted value calculates F1-scores for each class and then determines the weighted average. The weighting is based on the number of true instances for each class, making this metric a robust choice for scenarios where classes exhibit disparate sizes. This comprehensive evaluation encapsulates the balance between precision and recall in the context of varying class distributions.

4.8.5 Accuracy

Accuracy, a fundamental metric in model evaluation, measures the ratio of correctly predicted instances (both positive and negative) to the total number of instances. It is calculated using this formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

4.8.6 Area under the Curve (AUC)

AUC corresponds to the area under the Receiver Operating Characteristic (ROC) curve, which shows the performance of a binary classifier across various threshold settings. The ROC curve is constructed by plotting the True Positive Rate (TPR), as indicated in Equation 10, against the False Positive Rate (FPR), as shown in Equation 11. AUC ranges between 0 and 1, where an AUC of 0 indicates a model misidentifying all samples, an AUC of 1 signifies a model with 100% correct predictions, and an AUC of 0.5 represents a random predictor.

$$TPR = \frac{TP}{TP+FN} \tag{10}$$

$$FPR = \frac{FP}{TN+FP} \tag{11}$$

5 Results

In this section, we evaluate the performance of several machine learning models, including Random Forest, Gradient Boosting, XGBoost, and MLP, across patent and paper datasets. Key metrics such as precision, recall, F1-score, and AUC rates are presented to compare model effectiveness in detecting emerging technologies. The results highlight differences between patent and paper data and provide insights into the models' strengths, particularly in classifying rising and falling technological trends.

5.1 Patents

In Random Forest, we ran the model to see which hyperparameters work better. The best performance was achieved by having maximum depth of 12 trees and the number of estimators at 50. As the Random Forest serves as the baseline in the literature for this problem, we extensively report its results in Tables 3 and 4.

Table 3. Machine learning classification performance metrics for patent results

Metrics	Precision	Recall	F1-score	Support	Accuracy	Weighted F1
Falling	0.8525	0.6681	0.7491	476	-	-
Rising	0.5298	0.7639	0.6257	233	-	-
Overall	-	-	-	-	0.6996	0.7085

Table 4. Machine learning confusion matrix for patent results

Confusion Matrix	Predicted Falling	Predicted Rising
True Falling	318	158
True Rising	55	178

Table 3 shows the machine learning classification performance metrics for patent results, indicating a precision of 0.8525 and a recall of 0.6681 for the 'Falling' category, alongside a lower precision of 0.5298 but higher recall of 0.7639 for the 'Rising' category. This resulted in an overall accuracy of 0.6996 and a weighted F1 score of 0.7085, highlighting a balanced performance across categories. In Table 4, the confusion matrix for patent results is presented, showing the model's capability in accurately predicting 318

instances as 'True Falling' and 178 as 'True Rising,' but also indicating a notable number of instances (158) where 'True Falling' was incorrectly predicted as 'Rising,' and 55 instances of 'True Rising' being predicted as 'Falling.'

In Gradient Boosting, we ran the model to see which hyperparameters work better. The best performance was achieved by having maximum depth of 4 trees and the number of estimators at 100.

An exhaustive analysis was performed on the XGBoost model to pinpoint the ideal hyperparameters. Optimal performance was realized by limiting the tree depth to 12 and setting the estimator count to 100.

The best performance for the MLP model in the patent dataset was observed by having 2 layers, with 8 nodes in the first layer and 8 nodes in the second layer.

Our model accurately predicted that the prevalence of terms such as "convolutional neural networks" is falling, but terms such as "natural language processing," "reinforcement learning," and "object detection" are rising in patents in 2023.

The AUC rates and recall results of the tuned version of each method including Gradient Boosting, XGBoost, Random Forest, and MLP is in the Table 5.

Table 5. Machine learning paper results for the prediction of models

Model	AUC Test Data	Recall
Random Forest	76.58%	76.39%
Gradient Boosting	73.88%	61.80%
XGBoost	73.33%	48.93%
Multi-layer perceptron	76.73%	84.98%

As it is shown in the Table 5, the Random Forest model strikes a good balance between AUC and recall, landing at similar figures of 76.58% and 76.39%, respectively. Gradient Boosting, with a slightly lower AUC of 73.88%, falls behind more noticeably in recall, at 61.80%. XGBoost shows a comparable trend, achieving an AUC of 73.33% and a lower recall of 48.93%, which might suggest some difficulty in accurately identifying true positives. On the other hand, the MLP stands out with its high recall of 84.98% and solid

AUC of 76.73%, indicating its effectiveness in both prediction and accurately pinpointing relevant instances. These insights are quite useful for making informed choices when selecting models for specific predictive tasks in machine learning.

5.2 Papers

In the Random Forest model, an exploration of hyperparameters revealed that the highest performance was attained with a maximum depth of 9 trees and 64 estimators. As previously, since Random Forest is established as the baseline in the literature for this problem, we provide a detailed report of its results here.

Table 6. Machine learning classification performance metrics for paper results

Metrics	Precision	Recall	F1-score	Support	Accuracy	Weighted F1
Falling	0.8486	0.8608	0.8547	1537	-	-
Rising	0.7003	0.6793	0.6897	736	-	-
Overall	-	-	-	-	0.8020	0.7859

Table 7. Machine learning confusion matrix for paper results

Confusion Matrix	Predicted Falling	Predicted Rising
True Falling	1323	214
True Rising	236	500

Table 6 details the classification performance metrics for paper results, highlighting a high precision of 0.8486 and recall of 0.8608 for 'Falling', along with a precision of 0.7003 and recall of 0.6793 for 'Rising'. This leads to an overall accuracy of 0.8020 and a weighted F1 score of 0.7859, demonstrating the model's effectiveness. Additionally, Table 7 presents the confusion matrix for these results, showing that the model correctly predicted 1323 instances as 'True Falling' and 500 as 'True Rising', while misclassifying 214 as 'Predicted Rising' and 236 as 'Predicted Falling'. These results provide a comprehensive view of the model's performance, showcasing its strengths and areas for improvement in classifying paper results.

For Gradient Boosting, we conducted a comprehensive model assessment to identify optimal hyperparameters. The highest performance was achieved by setting the maximum depth to 8 trees and the number of estimators to 800.

We carried out an extensive evaluation of the XGBoost model to determine the best hyperparameters. The most effective results were obtained with a maximum tree depth of 16 and using 800 estimators.

The optimal configuration for the MLP model in the paper dataset was achieved with a 2-layer architecture, featuring 4 nodes in the first layer and 4 nodes in the second layer.

The AUC rates and recall results for the fine-tuned versions of each method, including Gradient Boosting, XGBoost, Random Forest, and MLP, can be found in Table 8.

Table 8. Machine learning paper results for the prediction of models

Model	AUC Test Data	Recall
Random Forest	77.01%	67.93%
Gradient Boosting	78.68%	71.74%
XGBoost	78.84%	71.74%
Multi-layer perceptron	86.51%	75.41%

As seen in Table 8, the Random Forest model showed a solid AUC rate of 77.01%, paired with a recall of 67.93%, indicating its reliable predictive capability. Gradient Boosting exhibited a higher AUC of 78.68% and a recall of 71.74%, suggesting improved accuracy in predictions. XGBoost reported similar performance, with a marginally better AUC of 78.84% and an identical recall to Gradient Boosting at 71.74%, highlighting its efficiency in certain scenarios. However, the MLP outperformed the other models with an outstanding AUC of 86.51% and a higher recall of 75.41%, showcasing its exceptional ability in both prediction accuracy and correctly identifying relevant cases.

5.3 Comparison of the AUC Results

In the literature, the AUC metric is the most important metric to compare different methods (Fawcett 2006). Consequently, we compiled the results into a single figure to facilitate a

more effective comparison of different models based on this metric.

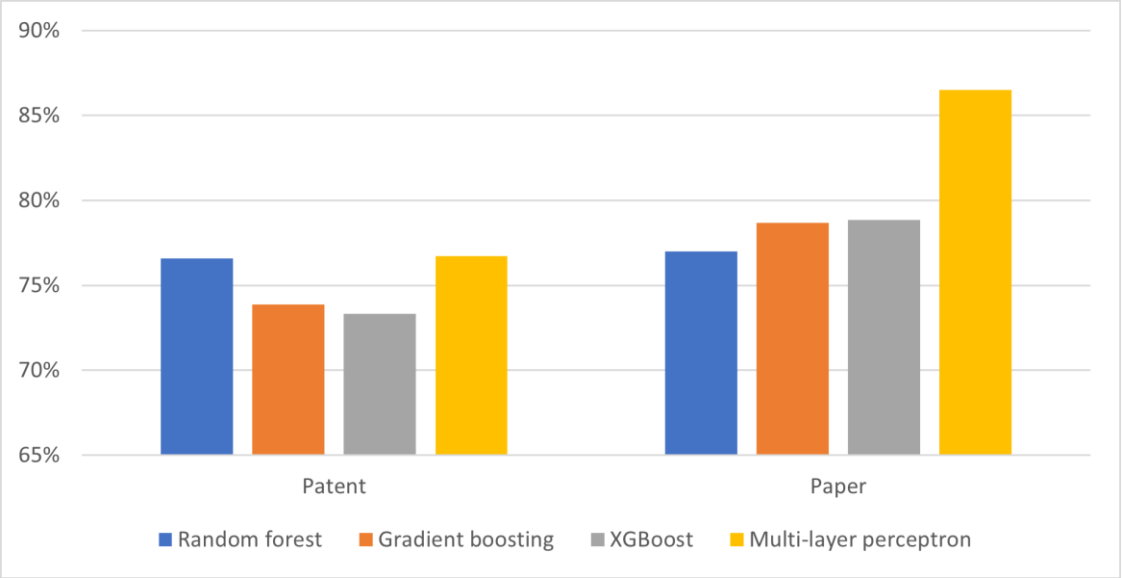

Figure 8. AUC results of different methods on patents and papers

Figure 8 shows the AUC results of different machine learning models applied to patents and papers. The models evaluated are Random Forest, Gradient Boosting, XGBoost, and MLP.

In the patent domain, MLP leads with the highest AUC result, which suggests it is the most effective among the models at distinguishing between the classes of interest in this context.

Gradient Boosting and XGBoost show similar performance for patents, with Gradient Boosting slightly edging out.

For papers, the MLP demonstrates a significant advantage with an AUC result approaching 86.51%, indicating it has a superior capability for classification tasks in this area compared to the others.

The AUC results for Random Forest, Gradient Boosting, and XGBoost are relatively lower for patents than for papers, showing these models may not perform as well with the patent data as they do with papers.

The AUC metric used here is a measure of a model's ability to distinguish between

positive and negative classes. Higher AUC values suggest better model performance. The marked difference in performance of the MLP for papers might imply that the MLP is better at capturing the nuances and complexities in the textual data often found in papers, compared to the more structured data of patents. This figure informs us about the relative strengths of each model in different domains and can guide decision-making in model selection for specific types of data.

5.4 Comparison of the Recall of Rising Technologies

As discussed before, the recall metric is one of the most important metrics inemerging technology detection. Therefore, we gathered the results in one figure to better compare different models in this metric.

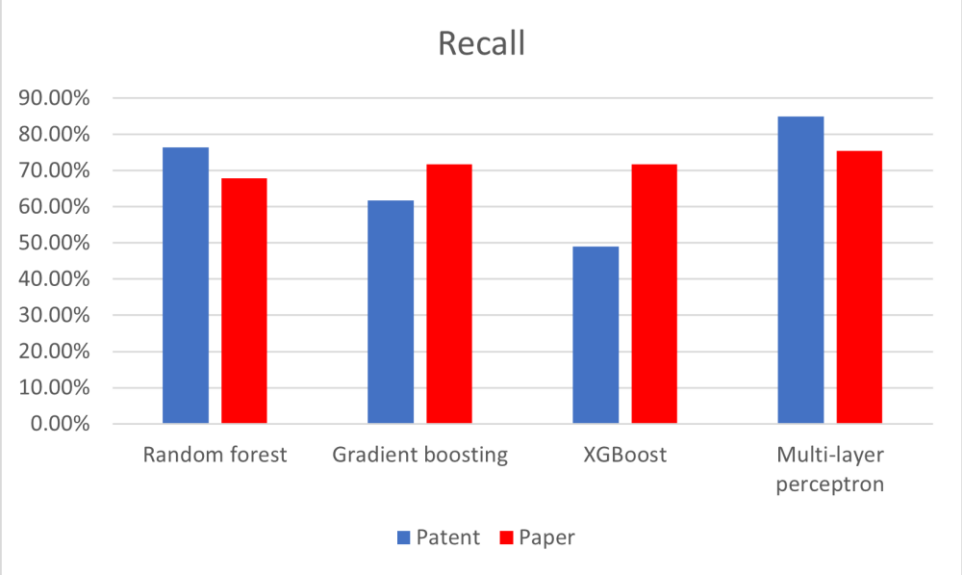

Figure 9. Recall results of different methods on patents and papers

Figure 9 presents a comparison of recall scores for rising technologies using different machine learning models: Random Forest, Gradient Boosting, XGBoost, and MLP, applied to two types of data, patents and papers.

The recall metric, crucial in the context of emerging technology detection, measures the model's ability to correctly identify all relevant instances. Below is the summary of the results:

- Random Forest shows a higher recall for patents compared to papers, indicating that it is more adept at identifying relevant technologies in patent data.
- Gradient Boosting has a balanced recall performance for both patents and papers, suggesting a consistent ability to recognize rising technologies across these domains.
- XGBoost appears to have a lower recall for patents but improves for papers, which may reflect an adaptation of the model to the different data structures or contents found in papers.
- The MLP shows the highest recall among all the models for papers and a strong performance for patents, suggesting its superior capability in recognizing relevant instances of emerging technologies, especially in the context of textual data analysis in papers.

These recall results provide insights into the efficacy of each model and help in determining which model might be best suited for analyzing specific types of data when it comes to detecting emerging technologies. The higher recall in papers for the MLP, in particular, underscores its potential usefulness in academic and research applications where paper data is prevalent.

6 Discussion

This chapter presents a discussion regarding the predicting of the sustainability of emerging technology terms, combining burst detection and deep learning.

Our results confirm that the burst detection method can successfully predict the future prevalence of the detected emerging technology terms with acceptable AUC accuracy rates for different data types with a 2-layer MLP model along with Random Forest, Gradient Boosting, and XGBoost models. By applying the trained MLP, Random Forest, Gradient Boosting, and XGBoost classifiers we predicted the future prevalence of emerging technology terms of unseen data using our data. We found that combining deep learning and burst detection can increase the AUC in comparison to other machine learning baseline methods. Also, by implementing the MLP on the dataset, we achieved higher AUC rates compared to previous research utilizing MACD applications for prediction.

6.1 MLP Dominance across both Datasets in Predictive Modeling

Here, baseline machine learning models, including Random Forest, Gradient Boosting and XGBoost, will be compared with a neural network classifier. The MLP classifier, leveraging MACD features, outperforms baseline models, demonstrating the effectiveness of deep learning in capturing complex relationships in time series data.

Between the four models applied across the two datasets—MLP, Random Forest, Gradient Boosting, and XGBoost—MLP consistently outperformed the baseline models in terms of AUC scores and recall. This consistent superiority underscores the efficacy of the MLP as a robust and versatile model capable of handling diverse datasets. Its inherent capacity to learn complex patterns and relationships has proven particularly advantageous in predicting emerging technologies and trends. The higher AUC scores further illustrate that the MLP consistently excelled in distinguishing between rising and falling prevalence instances, suggesting a higher degree of predictive accuracy and reliability. This demonstrates the MLP's superior ability to discern emerging trends, making it an invaluable tool in the study of technology forecasting.

Similarly, the higher recall rates for rising terms shows that MLP was more effective than other models in correctly identifying rising prevalence instances. This implies a greater

accuracy and dependability in the MLP's capability to pinpoint emerging trends comprehensively.

In essence, the prevalence of higher AUC and recall rates in the MLP models across different domains underscores its versatility and proficiency as a predictive tool in our study.

6.2 Application of MACD in Emerging Technology Detection

The application of MACD features for burstiness calculation and subsequent prediction of term emergence is a notable contribution. The approach aligns with the methodology of Tattershall et al. (2020), demonstrating consistency and reliability. The presented algorithm, considering burstiness levels and MACD values, shows a systematic way to predict future prevalence.

The AUC accuracy rates in different data sources, including patents and papers, further test the robustness of the proposed methodology. Also, having appropriate recall rates suggest that we can for use this model for identifying rising prevalence instances.

The integration of MACD into the emerging technology detection models along with machine learning techniques has produced notable AUC results, indicating the potential effectiveness of this technique in enhancing the discriminative power of the models. While XGBoost, Gradient Boosting, and Random Forest exhibit strong AUC values, the MLP stands out as particularly promising in capturing intricate trends. This is perhaps because have complex, non-linear relationships and interactions between features that may not be easily captured by more traditional models. The accuracy rates of rising and falling term prevalences are between 70% and 80%, indicating the potential of the MACD integration and emergence detection. Also, AUC rates with more than 70% in different models and datasets also show evidence of the potential of this model for future usage in emergence detection and prediction.

6.3 Effectiveness of the Proposed Emerging Technology Detection Method

The method developed in this thesis effectively addresses several critical limitations traditionally associated with the detection of emerging technologies. Firstly, the inherent subjectivity risk due to manual interventions is significantly mitigated by the automation

features of the burst detection and machine learning techniques. By reducing human involvement in the initial detection phases, the proposed method minimizes bias and enhances the objectivity of the detection process.

Regarding the lack of scalability, the integration of burst detection and machine learning models such as Random Forest, Gradient Boosting, XGBoost, and MLP allows the system to handle vast datasets efficiently. Burst detection is particularly effective with large volumes of data, enhancing the model's ability to accurately detect bursts and, in our case, the emergence of new technologies. This scalability is vital in adapting to the expansive and rapidly increasing volume of data concerning emerging technologies, making the method suitable for real-time and large-scale applications.

The lack of quantifiable metrics to evaluate the performance of traditional detection processes is remedied by the inclusion of various evaluation metrics in the proposed method. Metrics such as burstiness as a measure of emergence, along with AUC and recall metrics, offer clear, quantifiable indicators of performance. These metrics facilitate a systematic assessment of the method's effectiveness in identifying relevant emerging technologies and predicting their future relevance. Consequently, the detection process becomes not only more accurate but also verifiable.

Lastly, the issue of lack of predictability or low accuracy rates for future predictions is confronted by the advanced predictive capabilities of the MLP model. This model, in particular, has demonstrated high accuracy in forecasting the sustainability of emerging technologies, as evidenced by its superior AUC rate. This capability ensures that the predictions are reliable and actionable, enhancing the decision-making process for stakeholders involved in strategic planning and policy formulation regarding new technologies.

In summary, the proposed method not only addresses the key limitations found in traditional emerging technology detection methods but also provides a robust framework for continuous improvement and application in diverse technological fields.

7 Conclusion

Emerging technologies are unquestionably reshaping the world and various facets of our lifestyles, significantly influencing our future. In the dynamic and intricate landscape of emerging technologies, which attract substantial annual investments, early detection of emerging technologies proves to be both complex and costly. In addressing the challenges of early detection of emerging technologies, particularly in the realm of AI, this thesis has developed and validated an innovative methodology that leverages both burst detection and machine learning techniques.

The motivation for this approach stemmed from the significant limitations of traditional emergence detection methods, which often involve high degrees of manual intervention, subjectivity, and lack scalability and accurate predictions. Such limitations hinder their effectiveness in today's rapidly evolving technological landscape.

Our research objective was to create a more systematic, efficient, and objective method for the early detection of emerging technologies. The solution entailed to adapt a stock market-inspired algorithm for burst detection, enabling the identification and prediction of emerging technology trends in AI. This MACD process enabled us to provide machine learning features and inputs and also the output of the machine learning models. In the next step, by utilizing machine learning models—including Random Forest, Gradient Boosting, XGBoost, and the MLP classifier—we were able to forecast whether specific terms would gain or lose popularity over time.

By applying these methods to diverse data types, such as abstracts and titles from patents and research papers, we successfully predicted the future prominence of emerging technologies. Notably, the MLP model consistently outperformed other models, achieving the highest AUC and recall rates, with performance exceeding 75% across the two data sets.

This methodology not only reduced the need for manual intervention but also improved the scalability and provided acceptable accuracy rates of the emergence detection process. This research represents an important step toward combining burst detection with deep learning techniques to improve the detection and prediction of emerging technologies. Our findings

provide a strong foundation for further research and development in this area, contributing to more accurate and timely technology forecasting.

7.1 Limitations and future works

Admittedly, this research has its fair share of limitations. Firstly, results in this specific study may not generalize to other domains or scenarios, limiting the conclusions that can be drawn about its efficacy in broader applications, as it was only applied to one domain.

Secondly, even though noisy terms were managed by excluding irrelevant data and normalizing AI-related terms, the presence of new noisy terms underscores the inherent complexity and unpredictability in analyzing terms related to emerging technologies. There may still be noisy terms in our results reducing the quality of the final results. Improving methods for filtering noise, particularly in dynamic fields like AI, can increase the accuracy and reliability of predictions. Advanced preprocessing and data cleaning techniques might prove beneficial as noise reduction techniques.

Last but not least, applying MLP models despite having the potential in achieving higher accuracy rates might pose a certain risk. Given the complex nature of MLP and its capability to learn intricate patterns, there is a risk of overfitting, especially if the model is not adequately regularized or if the data is not sufficiently representative of broader trends. Also, the deep learning approaches, particularly MLP, may lack interpretability compared to simpler models. This can be a significant drawback in contexts where understanding the decision-making process is as important as the predictive accuracy. In future works, applying regularization techniques in models, especially deep learning ones like MLP, can help prevent overfitting, ensuring that the model generalizes well onto new, unseen data. Furthermore, we might choose to apply techniques like LIME (Local Interpretable Model-Agnostic Explanations) to increase interpretability, possibly thereby enhancing the understanding of the decision-making process.

References

Abercrombie, R. K., Udoeyop, A. W., & Schlicher, B. G. (2012). A study of scientometric methods to identify emerging technologies via modeling of milestones. *Scientometrics*, *91*(2), 327–342. https://doi.org/10.1007/s11192-011-0614-4

Adner, R., & Snow, D. (2010). Old technology responses to new technology threats: demand heterogeneity and technology retreats. *Industrial and Corporate Change*, *19*(5), 1655–1675. https://doi.org/10.1093/icc/dtq046

Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C., & Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics*, *15*(2), 101128. https://doi.org/10.1016/j.joi.2020.101128

Altuntas, S., Erdogan, Z., & Dereli, T. (2020). A clustering-based approach for the evaluation of candidate emerging technologies. *Scientometrics*, *124*(2), 1157–1177. https://doi.org/10.1007/s11192-020-03535-0

Ávila-Robinson, A., & Miyazaki, K. (2013). Dynamics of scientific knowledge bases as proxies for discerning technological emergence — The case of MEMS/NEMS technologies. *Technological Forecasting and Social Change*, *80*(6), 1071–1084. https://doi.org/10.1016/j.techfore.2012.07.012

Balili, C., Segev, A., & Lee, U. (2017). Tracking and predicting the evolution of research topics in scientific literature. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 1694–1697). Presented at the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA: IEEE. https://doi.org/10.1109/BigData.2017.8258108

Bengisu, M. (2003). Critical and emerging technologies in Materials, Manufacturing, and Industrial Engineering: A study for priority setting. *Scientometrics*, *58*(3), 473–487. https://doi.org/10.1023/B:SCIE.0000006875.61813.f6

Bettencourt, L. M. A., Kaiser, D. I., Kaur, J., Castillo-Chávez, C., & Wojick, D. E. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, *75*(3), 495–518. https://doi.org/10.1007/s11192-007-1888-4

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. (2017). A measure of staying power: Is the persistence of emergent concepts more significantly influenced by technical domain or scale? *Scientometrics*, *111*(3), 2077–2087. https://doi.org/10.1007/s11192-017-2342-x

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. (2018). An indicator of technical emergence. *Scientometrics*, *115*(1), 35–49. https://doi.org/10.1007/s11192-018-2654-5

Chen, T., Wang, Y., Fang, B., & Zheng, J. (2006). Detecting lasting and abrupt bursts in data streams using two-layered wavelet tree. In *Advanced Int'l Conference on Telecommunications and Int'l Conference on Internet and Web Applications and Services (AICT-ICIW'06)* (pp. 30–30). IEEE. https://ieeexplore.ieee.org/abstract/document/1602162/. Accessed 12 September 2024

Chen, Y.-H., Chen, C.-Y., & Lee, S.-C. (2011). Technology forecasting and patent strategy of hydrogen energy and fuel cell technologies. *International Journal of Hydrogen Energy*, *36*(12), 6957–6969. https://doi.org/10.1016/j.ijhydene.2011.03.063

Choi, Y., Park, S., & Lee, S. (2021). Identifying emerging technologies to envision a future innovation ecosystem: A machine learning approach to patent data. *Scientometrics*, *126*(7), 5431–5476. https://doi.org/10.1007/s11192-021-04001-1

Cozzens, S., Gatchair, S., Kang, J., Kim, K.-S., Lee, H. J., Ordóñez, G., & Porter, A.

(2010). Emerging technologies: quantitative identification and measurement. *Technology Analysis & Strategic Management*, *22*(3), 361–376. https://doi.org/10.1080/09537321003647396
Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, *73*(8), 981–1012. https://doi.org/10.1016/j.techfore.2006.04.004
Day, G. S., & Schoemaker, P. J. H. (2000). Avoiding the Pitfalls of Emerging Technologies. *California Management Review*, *42*(2), 8–33. https://doi.org/10.2307/41166030
de Rassenfosse, G., Dernis, H., Guellec, D., Picci, L., & van Pottelsberghe de la Potterie, B. (2013). The worldwide count of priority patents: A new indicator of inventive activity. *Research Policy*, *42*(3), 720–737. https://doi.org/10.1016/j.respol.2012.11.002
Ebadi, A., Auger, A., & Gauthier, Y. (2022). Detecting emerging technologies and their evolution using deep learning and weak signal analysis. *Journal of Informetrics*, *16*(4), 101344. https://doi.org/10.1016/j.joi.2022.101344
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
Furukawa, T., Mori, K., Arino, K., Hayashi, K., & Shirakawa, N. (2015). Identifying the evolutionary process of emerging technologies: A chronological network analysis of World Wide Web conference sessions. *Technological Forecasting and Social Change*, *91*, 280–294. https://doi.org/10.1016/j.techfore.2014.03.013
Garner, J., Carley, S., Porter, A. L., & Newman, N. C. (2017). Technological Emergence Indicators Using Emergence Scoring. In *2017 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 1–12). Presented at the 2017 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR: IEEE. https://doi.org/10.23919/PICMET.2017.8125288
Ghaemmaghami, A., Schiffauerova, A., & Ebadi, A. (2022). Which Keyword Extraction Method Performs Better for Emerging Technology Detection? In *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 613–618). Presented at the 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey: IEEE. https://doi.org/10.1109/ISMSIT56059.2022.9932656
Griol-Barres, I., Milla, S., Cebrián, A., Fan, H., & Millet, J. (2020). Detecting Weak Signals of the Future: A System Implementation Based on Text Mining and Natural Language Processing. *Sustainability*, *12*(19), 7848. https://doi.org/10.3390/su12197848
Griol-Barres, I., Milla, S., Cebrián, A., Mansoori, Y., & Millet, J. (2021). Variational Quantum Circuits for Machine Learning. An Application for the Detection of Weak Signals. *Applied Sciences*, *11*(14), 6427. https://doi.org/10.3390/app11146427
Guo, H., Weingart, S., & Börner, K. (2011). Mixed-indicators model for identifying emerging research areas. *Scientometrics*, *89*(1), 421–435. https://doi.org/10.1007/s11192-011-0433-7
He, D., & Parker, D. S. (2010). Topic dynamics: an alternative model of bursts in streams of topics. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 443–452). Presented at the KDD '10: The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC USA: ACM. https://doi.org/10.1145/1835804.1835862
Ho, J. C., Saw, E.-C., Lu, L. Y. Y., & Liu, J. S. (2014). Technological barriers and research

trends in fuel cell technologies: A citation network analysis. *Technological Forecasting and Social Change*, *82*, 66–79. https://doi.org/10.1016/j.techfore.2013.06.004
Iwami, S., Mori, J., Sakata, I., & Kajikawa, Y. (2014). Detection method of emerging leading papers using time transition. *Scientometrics*, *101*(2), 1515–1533. https://doi.org/10.1007/s11192-014-1380-x
Jang, W., Park, Y., & Seol, H. (2021). Identifying emerging technologies using expert opinions on the future: A topic modeling and fuzzy clustering approach. *Scientometrics*, *126*(8), 6505–6532. https://doi.org/10.1007/s11192-021-04024-8
Järvenpää, H. M., Mäkinen, S. J., & Seppänen, M. (2011). Patent and publishing activity sequence over a technology's life cycle. *Technological Forecasting and Social Change*, *78*(2), 283–293. https://doi.org/10.1016/j.techfore.2010.06.020
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. https://doi.org/10.1016/j.techfore.2016.08.020
Jun, S.-P. (2012). A comparative study of hype cycles among actors within the socio-technical system: With a focus on the case study of hybrid cars. *Technological Forecasting and Social Change*, *79*(8), 1413–1430. https://doi.org/10.1016/j.techfore.2012.04.019
Jun, S.-P., Yeom, J., & Son, J.-K. (2014). A study of the method using search traffic to analyze new technology adoption. *Technological Forecasting and Social Change*, *81*, 82–95. https://doi.org/10.1016/j.techfore.2013.02.007
Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change*, *117*, 228–237. https://doi.org/10.1016/j.techfore.2016.11.023
Kleinberg, J. (2002). Bursty and Hierarchical Structure in Streams. *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '02*, 91–101.
Lappas, T., Arai, B., Platakis, M., Kotsakos, D., & Gunopulos, D. (2009). On burstiness-aware search for document sequences. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 477–486). Presented at the KDD09: The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris France: ACM. https://doi.org/10.1145/1557019.1557075
Li, X., Xie, Q., Jiang, J., Zhou, Y., & Huang, L. (2019). Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. *Technological Forecasting and Social Change*, *146*, 687–705. https://doi.org/10.1016/j.techfore.2018.06.004
Liu, X., & Porter, A. L. (2020). A 3-dimensional analysis for evaluating technology emergence indicators. *Scientometrics*, *124*(1), 27–55. https://doi.org/10.1007/s11192-020-03432-6
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. https://doi.org/10.1016/j.techfore.2021.121159
Mane, K. K., & Börner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5287–5290. https://doi.org/10.1073/pnas.0307626100
Mejia, C., & Kajikawa, Y. (2020). Emerging topics in energy storage based on a large-

scale analysis of academic articles and patents. *Applied Energy*, *263*, 114625. https://doi.org/10.1016/j.apenergy.2020.114625
Murphy, J. (1999). *Technical Analysis of the Financial Markets*. Prentice Hall.
Nazarenko, A., Vishnevskiy, K., Meissner, D., & Daim, T. (2022). Applying digital technologies in technology roadmapping to overcome individual biased assessments. *Technovation*, *110*, 102364. https://doi.org/10.1016/j.technovation.2021.102364
Park, I., & Yoon, B. (2018). Identifying Promising Research Frontiers of Pattern Recognition through Bibliometric Analysis. *Sustainability*, *10*(11), 4055. https://doi.org/10.3390/su10114055
Porter, A. L., & Detampel, M. J. (1995). Technology opportunities analysis. *Technological Forecasting and Social Change*, *49*(3), 237–255. https://doi.org/10.1016/0040-1625(95)00022-3
Porter, A. L., Garner, J., Carley, S. F., & Newman, N. C. (2019). Emergence scoring to identify frontier R&D topics and key players. *Technological Forecasting and Social Change*, *146*, 628–643. https://doi.org/10.1016/j.techfore.2018.04.016
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. https://doi.org/10.3152/147154302781781001
Ranaei, S., Suominen, A., Porter, A., & Carley, S. (2020). Evaluating technological emergence using text analytics: two case technologies and three approaches. *Scientometrics*, *122*(1), 215–247. https://doi.org/10.1007/s11192-019-03275-w
Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, *44*(10), 1827–1843. https://doi.org/10.1016/j.respol.2015.06.006
Ruzzo, W. L., & Tompa, M. (1999). A Linear Time Algorithm for Finding All Maximal Scoring Subsequences. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, AAAI Press*, 234–241.
Santa Soriano, A., Lorenzo Álvarez, C., & Torres Valdés, R. M. (2018). Bibliometric analysis to identify an emerging research area: Public Relations Intelligence—a challenge to strengthen technological observatories in the network society. *Scientometrics*, *115*(3), 1591–1614. https://doi.org/10.1007/s11192-018-2651-8
Schiebel, E., Hörlesberger, M., Roche, I., François, C., & Besagni, D. (2010). An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices. *Scientometrics*, *83*(3), 765–781. https://doi.org/10.1007/s11192-009-0137-4
Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, *28*(11), 758–775. https://doi.org/10.1016/j.technovation.2008.03.009
Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology*, *60*(3), 571–580. https://doi.org/10.1002/asi.20994
Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., & Matsushima, K. (2011). Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological Forecasting and Social Change*, *78*(2), 274–282. https://doi.org/10.1016/j.techfore.2010.07.006
Tamura, K., & Kitakami, H. (2014). A new parallelization model for detecting temporal bursts in large-scale document streams on a multi-core CPU. In *2014 IEEE International*

Conference on Systems, Man, and Cybernetics (SMC) (pp. 519–524). Presented at the 2014 IEEE International Conference on Systems, Man and Cybernetics - SMC, San Diego, CA, USA: IEEE. https://doi.org/10.1109/SMC.2014.6973960
Tattershall, E., Nenadic, G., & Stevens, R. D. (2020). Detecting bursty terms in computer science research. *Scientometrics*, *122*(1), 681–699. https://doi.org/10.1007/s11192-019-03307-5
van Veen, B. L., & Ortt, J. R. (2021). Unifying weak signals definitions to improve construct understanding. *Futures*, *134*, 102837. https://doi.org/10.1016/j.futures.2021.102837
Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the Association for Information Science and Technology*, *69*(2), 290–304. https://doi.org/10.1002/asi.23930
Watts, R. J., & Porter, A. L. (1997). Innovation forecasting. *Technological forecasting and social change*, *56*(1), 25–47.
Watts, R. J., & Porter, A. L. (2003). R&D cluster quality measures and technology maturity. *Technological Forecasting and Social Change*, *70*(8), 735–758. https://doi.org/10.1016/S0040-1625(02)00355-4
Weismayer, C., & Pezenka, I. (2017). Identifying emerging research fields: a longitudinal latent semantic keyword analysis. *Scientometrics*, *113*(3), 1757–1785. https://doi.org/10.1007/s11192-017-2555-z
Wu, C.-C., & Leu, H.-J. (2014). Examining the trends of technological development in hydrogen energy using patent co-word map analysis. *International Journal of Hydrogen Energy*, *39*(33), 19262–19269. https://doi.org/10.1016/j.ijhydene.2014.05.006
Xu, H., Winnink, J., Yue, Z., Zhang, H., & Pang, H. (2021). Multidimensional Scientometric indicators for the detection of emerging research topics. *Technological Forecasting and Social Change*, *163*, 120490. https://doi.org/10.1016/j.techfore.2020.120490
Xu, S., Hao, L., An, X., Yang, G., & Wang, F. (2019). Emerging research topics detection with multiple machine learning models. *Journal of Informetrics*, *13*(4), 100983. https://doi.org/10.1016/j.joi.2019.100983
Xu, S., Hao, L., Yang, G., Lu, K., & An, X. (2021). A topic models based framework for detecting and forecasting emerging technologies. *Technological Forecasting and Social Change*, *162*, 120366. https://doi.org/10.1016/j.techfore.2020.120366
Yoon, J. (2012). Detecting weak signals for long-term business opportunities using text mining of Web news. *Expert Systems with Applications*, *39*(16), 12543–12550. https://doi.org/10.1016/j.eswa.2012.04.059
Yoon, J., & Kim, K. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics*, *90*(2), 445–461. https://doi.org/10.1007/s11192-011-0543-2
Zamani, M., Yalcin, H., Naeini, A. B., Zeba, G., & Daim, T. U. (2022). Developing metrics for emerging technologies: identification and assessment. *Technological Forecasting and Social Change*, *176*, 121456. https://doi.org/10.1016/j.techfore.2021.121456
Zhou, A., Qin, S., & Qian, W. (2005). Adaptively Detecting Aggregation Bursts in Data Streams. In L. Zhou, B. C. Ooi, & X. Meng (Eds.), *Database Systems for Advanced Applications* (Vol. 3453, pp. 435–446). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11408079_39

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. https://doi.org/10.1007/s11192-020-03351-6
Zhou, Y., Dong, F., Liu, Y., & Ran, L. (2021). A deep learning framework to early identify emerging technologies in large-scale outlier patents: an empirical study of CNC machine tool. *Scientometrics*, *126*(2), 969–994. https://doi.org/10.1007/s11192-020-03797-8
Zhou, Y., Lin, H., Liu, Y., & Ding, W. (2019). A novel method to identify emerging technologies using a semi-supervised topic clustering model: a case of 3D printing industry. *Scientometrics*, *120*(1), 167–185. https://doi.org/10.1007/s11192-019-03126-8
Zhu, Y., & Shasha, D. (2003). Efficient elastic burst detection in data streams. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 336–345). Presented at the KDD03: The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C.: ACM. https://doi.org/10.1145/956750.956789