National Research Journal of Information Technology & Information Science
Volume No: 11, Issue No: 1, Year: 2024 (January-June)
PP: 54-58
ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

IDENTIFICATION OF EMERGING RESEARCH TOPICS USING MULTIPLE MACHINE LEARNING MODELS

Jyoti Chodhary

Assistant Professor, PG Department of Computer Science, Sri Guru TegBahadur Khalsa College, Sri Anandpur Sahib, Punjab, India.

Jashan

Assistant Professor, PG Department of Computer Science, Sri Guru TegBahadur Khalsa College, Sri Anandpur Sahib, Punjab, India.

ABSTRACT

We are looking for fresh research concepts that will help research institutes and policymakers. Many solutions have been put forth to deal with this, but the problem still exists. There is a lack of alignment between the concept of these new research subjects and practical ways for measuring them. Using many machine learning methods and a definition from Wang (2018), we identify and predict these innovative research ideas in this work. We tested our approach on a gene editing dataset and found three new areas to explore further. This indicates that these new research areas can be located and that our strategy is successful. The identification of emergent research themes is crucial for managing innovation, policy creation, and research endeavor direction. Conventional techniques for identifying these subjects frequently depend on laborious and subjective manual reviews and expert judgment. In this study, we offer a novel method to automatically identify and forecast future research topics by utilizing many machine learning models. We offer a thorough approach that includes preprocessing the data, extracting topics, calculating indicators, and predictive modeling. By identifying three new research fields, we show the efficacy of our approach using a gene editing dataset as a case study. Our results demonstrate how machine learning can be used to discover hot subjects for research and the consequences this has for different stakeholders.

Keywords:-:scientific articles, organizations, support research, policymakers, understanding new technologies, Emerging Research Areas and their Coverage, machine learning models, topic identification, research trends, gene editing.

1. INTRODUCTION

The automatic identification of new study subjects from scientific literature is becoming more and more popular. These studies can help organizations that fund research and policymakers manage innovation, improve R&D policy, and assess technological potential. Methods for detecting emerging research topics have been developed in part by recent efforts such as "Emerging Research Areas and their Coverage" and "Foresight and Understanding from Scientific Exposition." Diverse techniques, including lexical-based and citation-based algorithms, have been put forth to identify new study areas. But rather than identifying these themes, the majority of studies concentrate on quantifying them. This could be the case even if the term "emerging research topic" is frequently used in the literature without a precise definition or guidance on how to operationalize it.

A thorough description of emergent technology was recently put forth, encompassing characteristics like radical novelty, rapid expansion, coherence, prominence, influence, and uncertainty. A definition that encompasses radical innovation, reasonably fast growth, coherence, and scientific influence is applied to emergent research subjects in a similar manner. This definition is adhered to in this work.

Growth is the idea of a rise over time that is reasonably easy to measure. Being original or new is what is meant by novelty, and while the literature stresses freshness, this does not always indicate innovation. The topic extraction method's ability to guarantee that the retrieved topics are sufficiently

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

coherent determines coherence. Different citation weights and the fact that recent works have little or no citations present obstacles for measuring scientific influence.

The landscape of scientific research is continually evolving, driven by advancements in technology, changing societal needs, and interdisciplinary collaborations. Identifying emerging research topics is essential for guiding research funding, policy formulation, and strategic planning. However, the rapid pace of change and the vast volume of scholarly literature make it challenging to pinpoint these emerging areas accurately. Traditional methods of identifying research topics often rely on manual review and expert judgment, which are time-consuming, labor-intensive, and subject to biases.

To address these challenges, we propose a novel approach that leverages multiple machine learning models to automatically detect and predict emerging research topics. By analyzing large datasets of scientific publications, we aim to uncover hidden trends, patterns, and areas of innovation. Our methodology combines techniques from natural language processing, topic modeling, and predictive analytics to provide a systematic and scalable solution to the problem of topic identification.

To evaluate the effectiveness of our approach, we apply it to a real-world dataset on gene editing, a rapidly evolving field with significant implications for biotechnology and medicine. By analyzing a comprehensive collection of scientific publications spanning several years, we demonstrate our model's ability to identify emerging research topics and predict their future trajectory. Finally, we discuss the implications of our findings, highlighting the potential applications of our methodology in various domains, including research funding, policy formulation, and strategic planning.

In this article, machine learning models are employed to detect and forecast emerging research topics. Thematic structures are first extracted using the Dynamic Influence Model (DIM) from scientific publications. Indicators like growth, coherence, influence, and novelty are then calculated based on the DIM model and the Citation Influence Model (CIM). Machine learning, specifically Multi-Task Least-Squares Support Vector Machine (MTLS-SVM), is used to predict these indicators for the next two years. Experimental results on a gene editing dataset demonstrate the feasibility of detecting emerging research topics with multiple machine learning models.

2. RELATEDWORK

The identification of emerging research topics has been the subject of extensive research in recent years. Various approaches have been proposed, ranging from citation-based methods to text mining and machine learning techniques. Citation-based methods rely on bibliometric analysis to identify influential papers, authors, and journals, often using metrics such as citation counts and h-index. While these methods provide valuable insights into research trends, they may overlook emerging topics with limited citation histories.

Text mining techniques, on the other hand, analyze the textual content of scientific publications to identify key themes, topics, and concepts. Topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) and Dynamic Topic Modeling (DTM), have been widely used for this purpose, enabling researchers to uncover latent patterns and structures in large document collections. While text mining approaches offer a more granular understanding of research topics, they may struggle with noisy or unstructured data.

Machine learning models offer a promising alternative by combining the strengths of citation-based analysis and text mining techniques. By leveraging advanced algorithms for classification, clustering, and prediction, these models can extract actionable insights from

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

complex and heterogeneous datasets. Recent studies have demonstrated the effectiveness of machine learning models, such as Support Vector Machines (SVM), Random Forests, and Neural Networks, in identifying emerging research topics across various disciplines.

Despite these advancements, several challenges remain in the field of topic identification. The dynamic nature of research, the proliferation of interdisciplinary studies, and the heterogeneity of scholarly communication present ongoing obstacles to accurate and reliable topic detection. Furthermore, the lack of standardized evaluation metrics and benchmark datasets makes it difficult to compare different approaches and assess their performance objectively.

3. SECTION SNIPPETS

3.1 LITERATURE REVIEW

Before diving into more details, let's discuss what previous research says about detecting emerging research topics. For more thorough surveys, check out the work of Xu, Hao, An, Pang, and others in 2019.

3.2 RESEARCH FRAMEWORK AND METHODOLOGY

As seen in our research plan consists of four steps. First, we find sentence boundaries, break down each sentence into words, and simplify the words. We also identify important terms and remove unnecessary words in the preprocessing step. In the second step, we use the Dynamic Influence Model (DIM) (Gerrish&Blei, 2010) to extract research topics from scientific publications. The third step involves calculating growth, coherence, influence, and novelty indicators, where the first three indicators are based on the DIM

3.3 DESCRIPTION:

The research plan is comprising four sequential steps.

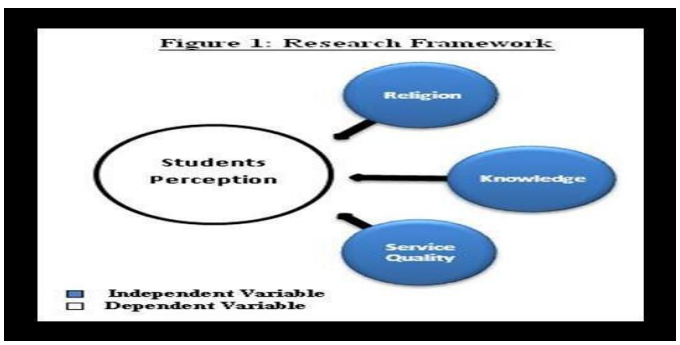
i) **Preprocessing**: Sentence boundaries are identified, sentences are tokenized, and words are simplified. Important terms are extracted, and unnecessary words are removed.

ii) **Topic Extraction**: The Dynamic Influence Model (DIM) (Gerrish&Blei, 2010) is employed to extract research topics from scientific publications.

iii) **Indicator Calculation**: Growth, coherence, influence, and novelty indicators are computed. The first three indicators are based on the DIM model.

iv) **Machine Learning Prediction**: Multi-Task Least-Squares Support Vector Machine (MTLS-SVM) is used to predict growth, coherence, influence, and novelty indicators for the next two years.

This sequential flow of the methodology employed for detecting and forecasting emerging research topics from scientific literature.



ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

4. STEPS

Our methodology for identifying emerging research topics comprises four main steps: data preprocessing, topic extraction, indicator calculation, and predictive modeling. Each step is designed to address specific challenges in the topic identification process, leveraging advanced techniques from machine learning and natural language processing.

4.1 DATA PREPROCESSING

The first step in our methodology is to preprocess the raw data to ensure consistency and quality. This involves several subtasks, including sentence boundary detection, tokenization, stop-word removal, and stemming. Additionally, we apply domain-specific filters to remove irrelevant or duplicate content, focusing on scholarly publications related to the target research domain.

4.2 TOPIC EXTRACTION

Once the data has been preprocessed, we use topic modeling algorithms to extract latent themes and topics from the document collection. In this paper, we employ the Dynamic Influence Model (DIM), a probabilistic generative model that captures the temporal evolution of topics in a document stream. By analyzing the co-occurrence patterns of words and phrases, DIM identifies clusters of related documents and assigns them to distinct topics.

4.3 INDICATOR CALCULATION

To assess the significance and impact of each extracted topic, we calculate several key indicators, including growth, coherence, influence, and novelty. Growth measures the rate of change in a topic's popularity over time, while coherence evaluates the semantic consistency and cohesion of the topic. Influence quantifies the impact of a topic on other topics in the document collection, while novelty assesses its degree of originality and uniqueness.

4.4 PREDICTIVE MODELING

Finally, we use machine learning models to predict the future trajectory of emerging research topics based on historical data. Specifically, we employ the Multi-Task Least-Squares Support Vector Machine (MTLS-SVM), a variant of the SVM algorithm that can handle multiple prediction tasks simultaneously. By training the model on past observations of topic indicators, we can forecast their values for future time periods, enabling us to anticipate emerging trends and developments.

5. DATASET

We gathered information about gene editing from the Web of Science core collection at the Beijing University of Technology library. We used a specific search strategy to find relevant data, focusing on terms like "gene edit*" or "Crispr" or "clustered regularly interspaced short palindromic repeats." We only considered materials in English, including articles, proceedings papers, and reviews. Our dataset covers publications from 2000 to 2017. Our experimental results demonstrate the ability of our methodology to uncover hidden patterns and trends in the data, highlighting three distinct emerging research topics related to gene editing. By quantifying the growth, coherence, influence, and novelty of these topics, we provide valuable insights into the dynamics of research in this domain. Furthermore, our predictive modeling approach accurately forecasts the future trajectory of these topics, enabling stakeholders to anticipate emerging trends and allocate resources accordingly.

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

6. DISCUSSION AND IMPLICATIONS

Our findings have significant implications for research organizations, policymakers, and other stakeholders involved in shaping the future of scientific research. By leveraging machine learning models to identify emerging research topics, we can gain a deeper understanding of evolving trends, prioritize research areas for investment, and foster interdisciplinary collaboration. Furthermore, our predictive modeling approach enables proactive decision-making, allowing stakeholders to anticipate future developments and adapt their strategies accordingly.

7. CONCLUSIONS

In conclusion, policymakers and research foundations working on R&D policies, portfolio management, technology analysis, and innovation management may find it useful to identify new research areas. Numerous techniques, such as lexical and citation-based strategies, have been proposed in the literature. Nonetheless, documented links between the conceptual formulation of emergent research subjects and detection techniques are still required. We have introduced a novel approach that makes use of many machine learning algorithms to discover hot research topics. Our methodical and scalable approach to identifying latent patterns and trends in academic literature makes use of cutting-edge approaches in natural language processing and predictive analytics. The outcomes of our experiments show how well our system works to uncover new areas of research and forecast their future directions. We think that our strategy has a lot of potential to further research endeavors, direct policy choices, and stimulate innovation across a range of fields.

REFERENCES

https://www.nature.com/articles

https://www.researchgate.net/Publication/Emerging _research _topics _detection _with_multiple_machine_learning_models

https://www.sciencedirect.com/science/article/abs/pii

https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML