Latent Dirichlet Allocation-Based Topic Mining Analysis of Educational Scientific Research Projects Based on 2360 NSF Education Projects

Jining Han [1], Geping Liu [1], Yuying Yang [1]

[1] *Faculty of Education, Southwest University, Chongqing, China*

Abstract **– The National Science Foundation has promoted the development of education in the U.S., and its establishment reflects the trend of education development. This study collects the data on 2360 NSF educational projects over the last three years to answer two research questions: What are the major research topics of NSF educational projects? What are the key projects doing? Through Latent Dirichlet allocation topic modelling, content analysis is carried out on the titles and abstracts of the 2360 projects, and eight topics are obtained from them, including top-notch innovative talent cultivation, STEM education for low-income groups, undergraduate education, vocational education, cutting-edge technology education, big data-driven technology, artificial intelligence, and teacher development.**

Keywords **- Latent Dirichlet allocation, NSF, educational projects, educational trends, implications.**

DOI: 10.18421/TEM122-32
https://doi.org/10.18421/TEM122-32

Corresponding author: Geping Liu,
Faculty of Education, Southwest University, Chongqing, China
Email: liugp2@swu.edu.cnl

Received: 19 February 2023.
Revised: 11 May 2023.
Accepted: 15 May 2023.
Published: 29 May 2023.

1. Introduction

Founded in 1950, the National Science Foundation (NSF) is a scientific research administration that advances the development of research and education through funding projects in basic science, engineering, and other fields. NSF's mission is to ensure the US's global leadership in science, engineering, and education by leading transformative research and excellence in science education to promote economic development, improve quality of life, and ensure national security [1]. By mining 2360 NSF education projects conducted in the past three years, this study found that research topics mainly focused on eight areas, comprising of STEM education, artificial intelligence, big data-driven technologies, and so forth. Since its establishment, it has funded the research of more than 200 Nobel Prize winners, supported tens of thousands of scientific research projects, and trained many of the best scientific and technological minds in the world [2], [3].

1.1. National Science Foundation Educational Projects

The research fields funded by NSF mainly include 12 aspects, together with computer science and engineering, education and human resources, mathematical and physical sciences, and behavioral and economic sciences, with the different research areas being covered by their respective divisions under NSF. Among them, the Education and Human Resources Bureau (EHR) mainly supports education at all stages from preschool to postgraduate and beyond, and it is responsible for promoting the cultivation of talents in emerging fields of science, engineering, and technology with the goal of building a diverse workforce of scientists, technicians, engineers, educators, and more.

EHR-funded research projects are closely integrated with education, and its four divisions have different goals: the Division of Research on Learning in Formal and Informal Settings (DRL) is committed to improving science and engineering education by promoting teaching research as well as development and evaluation of science and engineering in various learning environments; the Division of Undergraduate Education (DUE) focuses on improving the quality of undergraduate science and engineering education by improving curriculum, teaching, hardware, and evaluation; the Division of Graduate Education (DGE) mainly provides financial support for graduate students and develops innovative projects, and advocates high-quality, innovative graduate education in science and technology to train future leaders in this field; the Division of Human Resources Development (HRD) works to enhance participation in science and technology and encourages more teachers and learners to enter the field of science and technology.

1.2. Research Based on LDA Topic Modeling

LDA, based on the PLSA algorithm and originally proposed by Blei, is an unsupervised machine learning technique for identifying potential topic word information in document sets [4]. LDA topic model has proven to be well adapted in practical applications: for example, Wenjuan Wang and Jianxia Ma [5] used NSF-funded ocean acidification-related project applications and journal papers as research data for topic mining and evolution analysis, and the study found that LDA topic modeling can well mine the main research topics of these projects; Trivedi et al. [6] carried out topic modeling on the research trend of the COVID-19 pandemic, and found that LDA topic modeling can help identify heterogeneous topics related to the COVID-19 pandemic other than medical topic. Tan Chunhui et al. [7] constructed a sentiment analysis framework combining Baidu's sentiment analysis and LDA topic identification and studied the emotions of social media onlookers in privacy breaches. The results showed that topic modeling is helpful for further understanding of the causes of users' sentiment. Therefore, it is feasible to use the LDA model analysis to conduct data mining more systematically to analyze the funding trends and hot topics of NSF in education.

NSF-funded projects are the indicators of the development of scientific research and education in the US, and can be used as a reference for in-depth understanding of educational planning in the US. No research has yet used LDA topic modeling to analyze the topics and contents of NSF educational projects in depth.

Analyzing the textual data of the project declaration can summarize the current status and topics of NSF projects, as well as capture the key projects in the field. Therefore, this study analyzes the titles, keywords, and abstracts of NSF-funded educational project declarations in the past three years to uncover research topics and further analyze the hot issues and priorities of educational research with the help of LDA topic modeling. In this study the following two research questions are posed:

Research Question 1: What were the main topics of the projects funded by the NSF Education Department in the last three years?

Research Question 2: Among these themes, what are the representative key funded projects doing?

2. Research Methodology

This study employed the research method of LDA topic modeling, which first involved data collection and preprocessing. Next, the topic coherence and perplexity of the titles and abstracts of 2,360 NSF education projects were analyzed to determine the appropriate number of topic models. Finally, LDA topic modeling was performed.

2.1. Data Acquisition and Pre-processing

The research team queried the NSF EHR's project database for project information in 2020-2022 and as for October 2022, a total of 2360 project information was obtained. The main materials for topic modeling in this study are the titles, keywords, and abstracts of these 2,360 projects. The abstract can present the objectives, questions, main content, and findings of the study [8], [9]. Moreover, the combination of title, keywords and abstract condenses the essence of the study, and therefore the data is generally considered by the researchers to be suitable for conceptual review analysis [10], [11].

To improve the quality of the data for topic modeling, the research team pre-processed all the data. Firstly, the numbers, punctuation, symbols and stop words were deleted to ensure the accuracy of topic modeling. The stop words mainly consisted of two categories, one is modal particles, adverbs, prepositions and other words that usually have no clear meaning by themselves, such as, "I", "me", "or", "a", "they", etc.; the other category is common words that are frequently used in research grant applications, such as "study", "research", etc. Secondly, all plural words were converted into singular, for example, "teachers" has been converted into "teacher".

Thirdly, all capitalized initial letters were changed to lowercase, such as “Technology” has been changed into “technology”. Fourthly, words with multiple spellings were unified, such as “AI” has been replaced by “artificial intelligence”, “VR” by “virtual reality”.

2.2. LDA Topic Modeling

To generalize the topics of the 2360 projects, LDA topic modeling, a document topic generation model, is employed. As a conjugate prior probability distribution of a polynomial distribution, LDA can give the topics of a document in the form of a probability distribution. It is widely used in semantic mining, text clustering, topic detection, language evolution, etc [12], [13]. LDA unsupervised model can derive topic distributions and perform clustering by calculating the similarity, which has obvious advantages in terms of computational accuracy and clustering result [14]. Compared with latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA), non-negative matrix factorization (NNMF) and other topic modeling techniques, LDA has obvious advantages in accuracy and clustering result, allowing researchers to gain valuable insights and make data-driven decisions [15], [16].

In topic modeling, the first step is to use Python to calculate the topic perplexity and coherence to determine the optimal number of topics. The calculation of perplexity is to assign all words to an explicit topic, and then measure the modeling results by the likelihood of the model. Chang, Graber, Gerrish, Wang, & Blei [17] pointed out that the modeling results of perplexity cannot imitate human judgment very well, and it is difficult to understand and interpret. The calculation of coherence is based on the Normalized Pointwise Mutual Information (NPMI) formula, which assigns word pairs with high co-occurrence probability to a topic, thus making the modeling results highly comprehensible and interpretable [16], [18]. The NPMI formula is shown below.

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log \frac{p(w_i, w_j) + \varepsilon}{p(w_i) \cdot p(w_j)}}{-\log(p(w_i, w_j) + \varepsilon)}$$

p(w) is the probability of a word appearing in a given document, and p (w i ,w j) is the probability that these words appear together. NPMI calculates the statistical dependency between word pairs. If the value is higher, the topic model is better, that is, the higher the coherence, the better the modeling effect. This study applies topic coherence to determine the optimal number of topics, and it applies Python’s Gensim natural language processing library to implement LDA topic modeling, and finally employs pyLDAvis for visualization.

2.3. Key Funded Projects

The total funding for the 2360 projects established in the last three years was $1618.72 million dollars (as shown in Table 1). As shown in Figure 1, 4.83 % (114 projects) is funded over $2 million, 2.16% (51 projects) over $3 million, 0.7% (17 projects) over $ 4 million, and 0.3% (8 projects) over $5 million. The research team first generated topics based on LDA topic modeling, and then selected representative projects above $2 million from each topic to provide an in-depth understanding of each topic on the one hand, and to analyze key funded projects on the other.

Table 1. Number and Funds of Educational Projects in 2020-2022

	2020	**2021**	**2022**	**total**	**funds**
DGE	75	92	64	231	224384139
DRL	124	117	94	335	238510462
DUE	415	463	501	1379	955685601
HRD	119	151	145	415	200148058
total	733	823	804	2360	1618728260

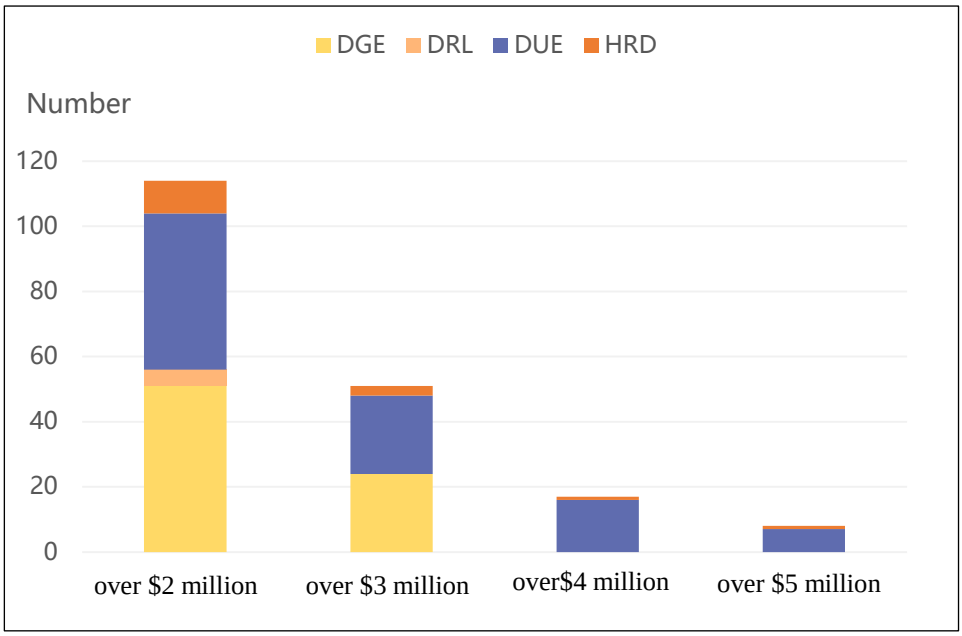

Figure 1. Number of Key Educational Projects in 2020-2022

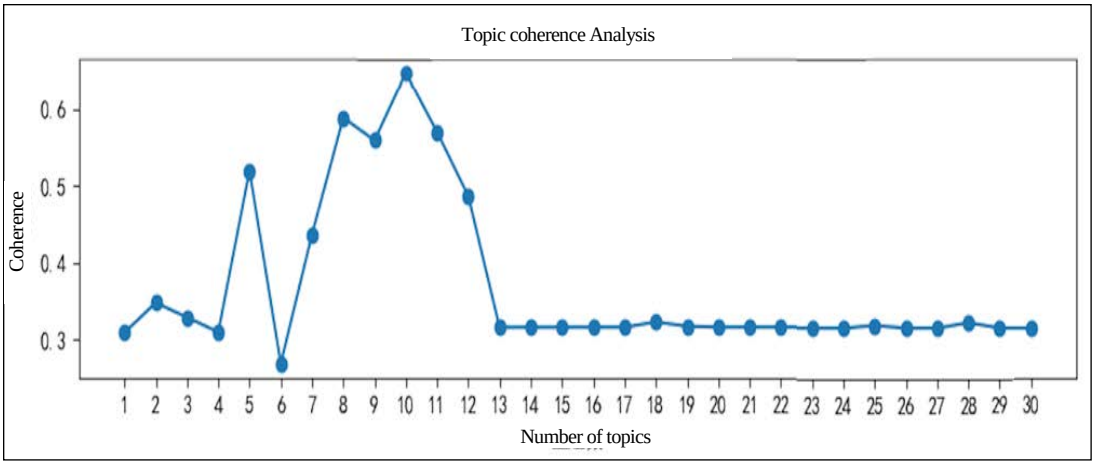

Figure 2. Topic Coherence Analysis

3. Research Results

In this study, a topic coherence analysis is conducted on the titles, keywords, and abstracts of the 2360 NSF educational projects. The results of the analysis show (see Figure 2) that the highest coherence is achieved when the number of topics is 10, which can achieve the best modeling effect, so 10 is chosen as the optimal number of topics.

The research team applied 10 topics for LDA topic modeling, so as to more comprehensively and clearly dig out the most concerned research topics of NSF educational funds in the past three years (as shown in Table 2). The table lists the high-frequency words in each topic, and the values after each vocabulary indicate the weights. The visualization data was then generated by using pyLDAvis (as shown in Figure 3). The circle area indicates the importance of each topic in the entire corpus, the distance between the centers of the circles indicates the similarity between the topics, and the histogram on the right lists the top 30 most relevant words for each topic. From the output data, it can be seen that topic 9 has a high overlap rate with topics 1 and 4, and is relatively close to the center of the circle of topic 10. The repetition rate of these topic words is high, which means that these topics have a high degree of similarity, so the research team deleted the topic 9 and topic 10 output by LDA topic modeling. The research team summarized eight important topics according to the results of topic modeling and the names and abstracts of projects that classified into the topics.

Topic 1: Cultivation of top-notch innovative talents in higher education

The keywords of topic 1 extracted by LDA topic modeling include "STEM", "support", "academic", "engineering", "faculties", "equity", "advanced", "university", "research", and "develop".

By analyzing the extracted key words and the projects corresponding to topic 1, it is found that the projects in topic 1 mainly focus on the cultivation of top-notch innovative talents in science and engineering in universities and scientific research institutions. In the context of this study, STEM not only refers to the combination of the four disciplines of science, technology, engineering, and mathematics. What is commonly referred to as STEM education or maker education also represents a cluster of disciplines that encompasses a wide range of science and engineering majors. For example, Dartmouth College's "Transformative Research and Graduate Education in Sensor Science, Technology and Innovation" project, which received a $3million grant in 2021, aims to build an innovative talent training system that integrates student-centered STEM education with sensor professional courses. It focuses on developing students' innovation and entrepreneurship, and training 129 interdisciplinary masters and doctors to become global leaders in sensor technology, thereby fostering innovation in areas such as health medicine, remote sensing, and new energy. Michigan State University's "Using Data Innovation to Enable Multi-Scale Prediction Models for STEM Education" project, which received a $2.96 million grant in 2022, aims to create an innovative, evidence-based talent training system that meets the needs of the STEM workforce. By building a multi-scale prediction model for the future STEM workforce and developing STEM core courses and targeted intensive courses based on the model, the project will prepare 100 doctoral students who will go on to careers in STEM research and education in academia, industry or national laboratories after their graduation.

It is worth noting that the word "equity" appears in the cultivation of top-notch innovative talents. After excavating key projects, we found that "equity" is mainly reflected in this topic by: (1) promoting equity through top-notch innovative talents, and (2) the equity of talent selection. Massachusetts Amherst's "Enhancing Equity and Resilience in Sustainable Energy Development" project, which received $ 3.05 million grant in 2022, brings together students from diverse educational and socio-economic backgrounds to train 125 masters and doctors by offering STEM-related courses. The goal is to integrate the concept of equity into the research and education of energy transition to achieve equity driven innovation, so that students can both equitably benefit society and adapt to future climate and energy development to contribute to the global low-carbon society. The University of Minnesota, Twin Cities' "Cultural Diversity and Social Equity as Drivers of Water and Materials Cycles" project, which received $3 million grant in 2022, emphasizes that advancing resource cycles through technology has had little effect and needs to start focusing on cultural perspectives, economic backgrounds and policy frameworks.

It is necessary to select 300 students from different cultural and economic backgrounds in a multi-dimensional and fair manner, so that they can use new technologies in the future and take into account local cultural and economic conditions to construct appropriate approaches for the benefit of society.

Topic 2: Supporting STEM education among low-income groups

The keywords of topic 2 extracted by LDA topic modeling include "low income", "support", "STEM", "demand", "engineering", "finance", "retention", "graduation", "improvement", "scholarship". From this we can see that the projects of topic 2 revolve around the use of incentives such as scholarships and employment need orientation to encourage low-income students to continue their studies and complete STEM-related studies. The University of California, San Diego's "Helping Low-income Students Succeed in Engineering through Practice" project, which received $5 million grant in 2022, plans to help 185 low-income undergraduate majoring Engineering complete their studies over five years. The researchers will evaluate this funding approach and ultimately form an educational model framework that supports students from low-income families. The University of Augsburg's "STEM Professional Enhancement for Minnesota Low-income Students" project has been given $3.07 million in 2020. It aims to explore how to support the education and employment of students from low-income families. This research will recruit 195 low-income students to pursue STEM majors by means of transferring majors or schools, provide these students with a series of STEM courses, and ultimately support them into employment through inter-institutional partnerships. Based on five years of tracking data, the researchers will evaluate the key factors affecting the academic and employment of these students.

Topic 3: Improving the Teaching Quality of Undergraduate STEM Education

The keywords of topic 3 extracted by LDA topic modeling include "STEM", "support", "learning", "courses", "science", "influence", "undergraduate", "college", "Mathematics", "community". The projects in topic 3 focus on the teaching of STEM at undergraduate level, which includes many elements such as support models, curriculum settings, influencing factors, and teaching methods and more. East Stroudsburg University's "Supporting Students to Earn Undergraduate Degrees in Mathematics and Science through Institutional Partnerships" program received $ 4.98 million grant in 2021.

The project aims to integrate the advantageous disciplines of multiple local educational institutions, and through financial incentives, collaborative training, peer assistance, online teaching, and personalized development plans, etc., to change the way of teaching and talent training of undergraduate mathematics and science.

In five years, 135 undergraduates will be trained in this way to ensure that 90% of the students graduate with good grades, and finally form a model project that can be replicated in other regions. "STEM Teaching at Undergraduate Level Intervention" project of California State University Channel Islands, which received $2.5 million grant in 2020, aims to improve

Table 2. LDA Topic Modeling Output Results

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
STEM	low-income	STEM	STEM	STEM	STEM	AI	teacher	STEM	STEM
0.012	0.018	0.031	0.011	0.022	0.020	0.008	0.040	0.009	0.022
support	support	support	support	support	faculty	support	STEM	learn	support
0.011	0.016	0.015	0.010	0.017	0.017	0.008	0.019	0.009	0.012
academic	STEM	learn	learn	development	support	learn	science	support	learn
0.010	0.014	0.010	0.009	0.015	0.013	0.007	0.013	0.009	0.009
engineering	need	course	development	learn	college	university	support	impact	undergraduate
0.009	0.013	0.010	0.006	0.015	0.009	0.007	0.011	0.007	0.007
faculty	engineering	science	community	impact	community	science	school	development	impact
0.009	0.012	0.009	0.006	0.009	0.007	0.007	0.009	0.007	0.007
equity	financial	impact	university	science	institution	healthcare	preservice	college	development
0.009	0.011	0.009	0.005	0.008	0.007	0.006	0.008	0.006	0.006
advance	retention	undergraduate	workforce	engineering	impact	development	development	science	university
0.008	0.011	0.008	0.005	0.006	0.006	0.006	0.008	0.005	0.006
university	graduation	college	science	undergraduate	new	STEM	university	community	course
0.008	0.010	0.008	0.005	0.006	0.006	0.006	0.008	0.005	0.006
institution	increase	mathematics	technology	institution	data	impact	learn	faculty	projects
0.007	0.009	0.008	0.005	0.006	0.006	0.005	0.007	0.005	0.006
development	scholarships	community	course	faculty	university	merit	teaching	undergraduate	faculty
0.006	0.009	0.008	0.005	0.006	0.006	0.005	0.006	0.005	0.006

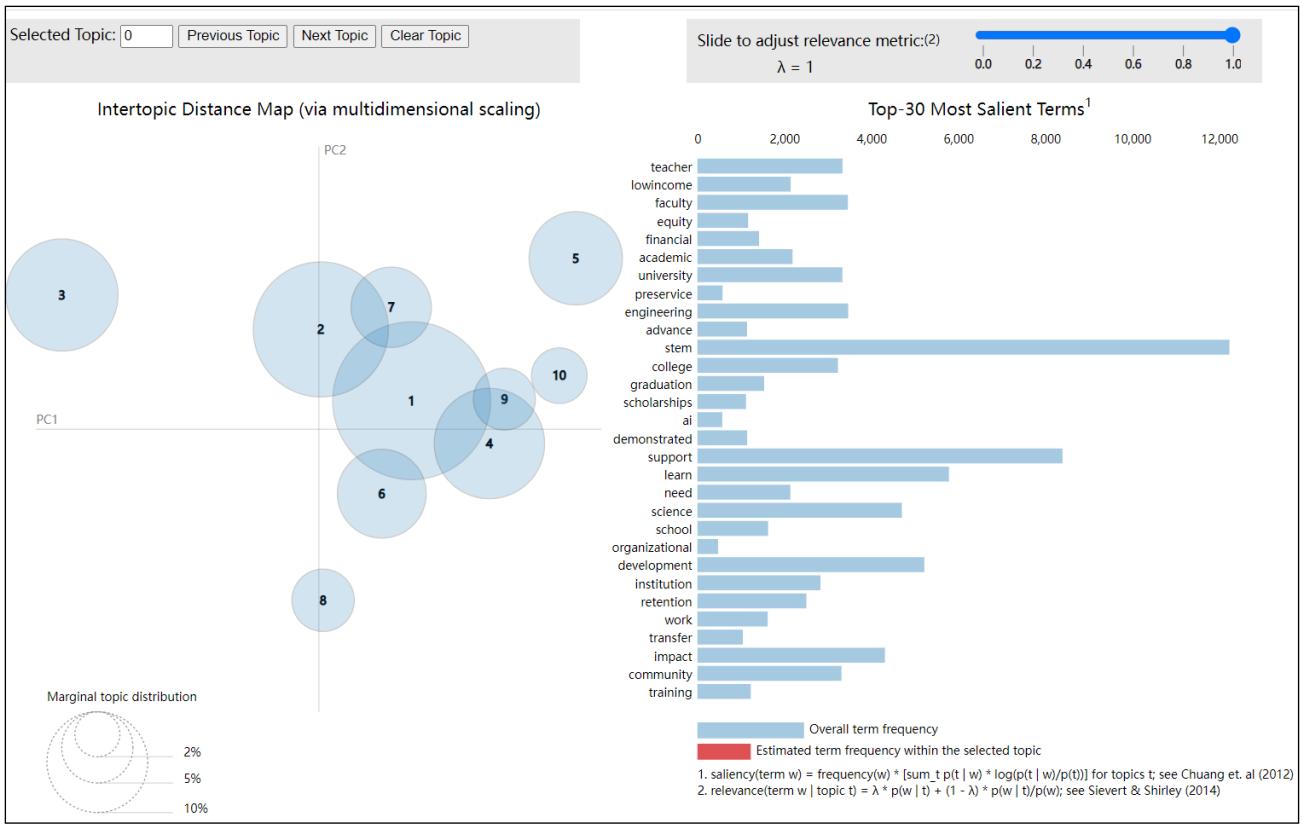

Figure 3. Visualization of LDA Data Results

STEM teaching at undergraduate level through a series of interventions, including offering summer intensive courses, providing psychological counseling for students, arranging research mentors for undergraduates, encouraging undergraduates to participate in scientific research and work closely with the community. The researchers aim to test whether these interventions can help STEM undergraduates reduce academic anxiety, improve academic performance and graduation rate, and finally form an intervention model for the high-quality development of STEM education at undergraduate level.

Topic 4: STEM education in community college that is aligned with the labor market

The keywords of topic 4 extracted by LDA topic modeling include "STEM", "support", "learning", "development", "community school", "university", "workforce", "science", "technology", "course". In the US education system, community colleges, also known as technical colleges, two-year colleges or city colleges, mainly attract and accept students from the local community, and prepare them for employment through partnership with local businesses. Topic 4 reflects NSF's focus on preparing STEM students at community colleges to meet the needs of local businesses and the labor market. The American Association of Physics Teachers' "Physics Education for Diverse Populations: Advancing Physics Teaching at Two-Year Colleges" project received $ 2.92 million grant. It aims to fill the gap in physics teaching in community colleges. The specific content includes systematizing and institutionalizing physics teaching in community colleges, training physics teachers, practicing new physics teaching methods, and forming a typical example of STEM education in community colleges. The project is expected to have a positive impact on 110,000 students within five years and promote the development of STEM education in community colleges. Pasadena City College's "Micro-Nano Technology Education Center" project, which received $9.02 million grant in 2020, will unite four community colleges to train technicians engaged in the micro-nano field, and provide an adequate skilled workforce for the manufacturing of micro-nano products nationwide. "National Next Generation Manufacturing Center" project of Tunixs Community College, which received $7.5 million grant in 2020, aims to cultivate a skilled workforce in advanced manufacturing. It plans to cultivate technical teams that can engage in processing, design, manufacturing, supply chain, logistics, and quality control to meet the needs of advanced manufacturing through the integration of online educational resources, distance learning, internships and apprenticeships.

Topic 5: Education on cutting-edge technologies in STEM

The keywords of topic 5 extracted by LDA topic modeling include "STEM", "support", "development", "learning", "impact", "science", "engineering", "undergraduate", "scientific research institutions", "faculties". Although the keywords in topic 5 are close to those in other topics, as shown in Figure 3, the distance between the centers of the circles indicates the similarity between topics, and the center of topic 5 is farther away from other topics. After digging further into the projects in topic 5, we find that topic 5 focuses on recent cutting-edge technologies in STEM, such as quantum science education and cybersecurity education. The University of Washington's "Interdisciplinary Linking Quantum Sensing Technology: A Training Program Integrating Quantum Science and Engineering" project, which received $2.99 million grant in 2022, plans to offer courses related to quantum science through an interdisciplinary approach. It is expected to train 100 students as engineers in fields such as astrophysics, materials, medical imaging and more. "Comprehensive Training in Quantum Assembly and Technology" project of the University of California, Santa Barbara received $2.98 million grant in 2022. It will provide new courses on quantum science, and it is expected to support 30 students to become the next generation of quantum scientists. Whatcom Community College's "National Cybersecurity Training and Education Center", approved for $7.49 million in 2021, will develop and deploy leading cybersecurity courses, cultivate cybersecurity majors, and widely apply cybersecurity education to universities, primary and secondary schools under the guidance of research universities such as California State University and Embry-Riddle Aeronautical University.

Topic 6: Big data education in STEM subjects

The keywords of topic 6 extracted by LDA topic modeling include "STEM", "faculties", "support", "college", "community", "scientific research institution", "impact", "innovation", "data", "university". Big data education is still a major focus of educational projects in the last three years. Topic 6 shows that community colleges, comprehensive universities and scientific research institutions are all paying attention to big data education in STEM.

Michigan State University's "Data-Driven Economic Innovation" project, which received $ 4.98 million grant in 2022, will attract more students to take STEM courses and is expected to train 157 undergraduates majoring in data science, with the aim of enhancing students' data mining and analysis skills to address new challenges in a data-driven economy. Iowa State University's "Reliable Data-Driven Decision-Making" project has been approved for $2.99 million in 2022. This project will first build a "Reliable Data-Driven Discovery Framework" that will lead to methodologies and application tools to address noise in data, limited data sets, uncertainty in predictions, interpretability of machine learning and other issues, thus forming a risk inspection system. In the process, it is expected to train 100 students in STEM. These students will complete the courses and work in the project to improve their literacy in data analysis and risk warning.

Topic 7: Artificial Intelligence in education

The keywords extracted by LDA topic modeling for topic 7 include "artificial intelligence", "support", "learning", "university", "science", "health care", "development", "STEM", "impact", "advantage". Topic 7 mainly focuses on the application of AI in education. San Jose State University's "Expanding Participation in Undergraduate Artificial Intelligence Education Though Situational Learning" project, which received $ 2.63 million grant in 2022, aims to meet national needs by improving AI education at universities. The research team will develop AI learning modules, offer relevant courses, and attract more students to choose AI as a major so that students can solve social problems based on the theory and application of AI. The University of Missouri-Kansas' "Interdisciplinary Graduate Training: Mitigating the Alcohol and Drug Abuse Crisis through Artificial Intelligence and Network Sensing" project has been approved for $2.99 million in 2022. It will combine courses in AI, communication networks, sensor technology, and health care provide a unique and comprehensive education for 120 graduate students. It analyzes consumers' health history, personal background and behavior data through AI and network sensing technology so as to deal with the nation's growing alcohol and drug abuse crisis.

Topic 8: Supporting the professional development of STEM teachers

The keywords extracted by LDA topic modeling for topic 8 include "teacher", "STEM", "science", "support", "school", "pre-service teacher", "development", "university", "learning", "teaching".

Topic 8 focuses on the cultivation and professional development of teachers in STEM.

The American Association for the Advancement of Science's "STEM Teacher Preparation Program for High-Demand School Districts" project, approved for $ 3.48 million in 2020, aims to support STEM undergraduates to become STEM teachers in primary and secondary schools, and to improve teachers' abilities and expand their professional knowledge through a series of trainings, and help teachers build teaching and research communities. This will ensure the long-term, stable, and high-quality development of STEM teachers in school districts with high demand for STEM education. With a $ 2.52 million grant in 2022, Cal Poly's "Training Successful Pre-service Teachers" project aims to meet national needs for high-quality STEM teachers with practical experience before they start their career. The project team expects to provide 120 nine-week internships for local pre-service STEM teachers within three years, and through tracking research to examine the development status of pre-service teachers, so as to provide suggestions for their future teaching practice.

4. Discussion and Implication

Based on the results of data analysis using LDA topic modeling, this study identified four areas that warrant further investigation. Therefore, this study conducted an in-depth discussion on the following topics: the urgent need for targeted talent, improving economic development through education, pursuing new technologies, and integrating research and teaching.

4.1. The urgent need for targeted talents: multi-level and all-round training of STEM professionals

In the NSF project declaration, it is repeatedly mentioned that "meeting the nation's need for STEM professionals", which is mainly reflected in two aspects, one is the need for high-tech innovators, and the other is the skilled labor force for advanced manufacturing. Some projects explicitly state that the nation has a strong need for well-educated scientists, mathematicians, and engineers to ensure the growth of high-tech innovation and research in the US. On the other hand, the nation has a growing demand for skilled labor in advanced manufacturing, with a 2018 study by Deloitte and the Manufacturing Institute reporting that the US will need to fill more than 4.6 million advanced manufacturing jobs, and that more than half of these jobs will be vacant due to the gap of workers' skills. This dilemma is inseparable from the US government policies, many of which have led to a talent shortage in the high-tech market.

For example, the International Traffic in Arms Regulations (ITAR) law requires high-tech enterprises to hire only US citizens in some high-tech fields, such as aerospace, navigation systems, and high-tech components. In addition, although the two-year graduation internship period has been extended for STEM graduates, it is only an immediate solution and cannot guarantee a steady supply of innovation and labor force. As a result, federal policy makers have paid close attention to the US STEM education system in recent years and are eager to carry out educational work in STEM. NSF is a key component of supporting the STEM education system, providing the largest number of funds and funded projects for STEM education. The last three years of projects have shown a divergence in the pathways for training talent in STEM: a trend has been formed that "top-notch innovative talents belong to R&D, undergraduate talents belong to application, and skilled labor talents belong to manufacturing". The project funds are used in a hierarchical and precise way. For research-oriented universities and research institutions, it mainly focuses on the research and development of high-precision technology and the cultivation of top-notch talents. Funds also are used in undergraduate education to improve the literacy of application engineers and operators. At the same time, a large amount of funds is invested in STEM education at community college to cultivate skilled workers for advanced manufacturing in a short period of time.

4.2. Improving the economic development through education: to enhance the socio-economic situation of families through STEM education

The US has been faced with the huge challenges of multiracial integration and poverty. For example, in Chicago, Illinois, the 2020 census report pointed out that one in ten residents is in extreme poverty, with these residents earning less than 50% of the national poverty line per capita income and one in four children living in poverty. The report further stated that the biggest problem facing Chicago is the low quality of education for low-income groups, which leads to the intergenerational transmission of poverty. New York City is also facing the same problem. According to the 2018 New York City government survey report, the poverty rate in New York City is close to 20%. What's more serious is that 41.3% of the population is hovering at the poverty line [19]. In addition to the big cities, rural areas also have the same problem. In 2019 the Tennessee government reported that its poverty rate also reached 13.9 %.

This has created a vicious cycle in which some low-income families in the US do not value education, and their children cannot successfully complete their studies and are unable to prepare for the labor market, leading to a generation that continues to be poor. In the long term, it has caused a severe shortage of skilled labor in the US manufacturing sector. According to the U.S. Census Bureau, the number of poor people living in poverty is 37.9 million, accounting for 11.6% of the total population. It is worth noting that the poverty rate for those under 18 is as high as 15.3%, the poverty rate for those with a bachelor's degree or higher is only 4.1%, and the poverty rate for those with full-time jobs is only 1.8%. NSF funds in the past three years have paid great attention to encouraging students from low-income groups to enter STEM subjects, and supported them to complete their studies through significant funding. According to the eight topics in the data results, the three topics of STEM education for low-income groups, STEM education at community colleges, and STEM education at the undergraduate level are all targeted at improving the education quality of students in low-income groups, which has a huge population base. If STEM education is used to cultivate advanced manufacturing production labor skills, it can solve the labor shortage on the one hand, and increase the employment rate of low-income families on the other, thereby alleviating the current poverty challenge as well as addressing the vicious circle of intergenerational transmission of poverty.

4.3. Pursuing new technologies: continue to focus on advanced technological fields such as quantum science, artificial intelligence, and big data

Artificial intelligence and big data still have a place in the projects of the last three years. The data results of this study echo the ecological picture of AI education in the US drawn by Xiaoqing Gu and other scholars in 2021. It is reflected in key projects that AI empowers learning, enhances learning ability, and promotes the equity and sustainable development of education. At the same time, we also found that in recent years, AI education has paid extra attention to combining AI with medical and healthcare education. AI education is used to solve practical problems in society. In addition to AI and big data, NSF educational projects are also pursuing the education of the latest technologies, especially quantum science education. The progress of quantum science requires the combination of knowledge and technology in multiple fields, and it also needs talents to be cultivated in different fields.

If we want to achieve the "quantum leap", we must break the current dominant mode of postgraduate education, where different departments teach in isolation, and carry out cross-faculty and cross-professional talent training to form a new top-notch innovative talent training model. The development of quantum science in the future needs to be driven by a new generation of quantum scientists who are receiving training in scientific research universities and institutions, making interdisciplinary quantum science education a national priority. To take the lead in quantum science, it is imperative to focus on the cultivation of talents in quantum science.

4.4. Integrating research and teaching: closely combining talent training and scientific research projects

Through the analysis of the key projects, we find that scientific research projects are closely integrated with talent training, and the deep integration between the two mainly presents in two aspects: one is that some scientific research projects are aimed at talent training; the other is that students are involved in scientific research projects and talent training is carried out in the process of research. In the first category, we can see that scientific research projects attach great importance to talent training, with most of them specifying the specific number of students to be trained within a certain period of time and also indicating the future development of these students. The second category of projects uses the method of cultivating talents in the process of scientific research to organically integrate scientific research innovation and talent training. This is made possible by the flexible credit system of universities in the US and the great autonomy of teachers in setting up courses. The faculties with scientific research funds can independently offer courses or invite other experts to offer courses based on research needs and encourage students to take these courses. Students have the opportunity to practice in scientific research projects while learning the knowledge, so that they can apply what they have learned. The bi-directional relationship has promoted the long-term, continuous, stable and high-quality development of scientific research and teaching, as scientific research projects need talents, and talents also need to grow and progress in scientific research.

Acknowledgements

This study was financially supported by the iFLYTEK AIED Qingteng project "Research on AIED promotes students' math computing and intuitive imaginational literacy" under Grant "01O20220509EBGZ003", the Center for Language Education and Cooperation project

"Research on AIED assists language learning" under Grant "22YH92C", "2019 Guangdong Social Development Science and Technology Collaborative Innovation System Construction Project" under Grant "2019B110210001", and the Southwest University Innovative Research 2035 project under Grant "SWUPilotPlan002".

References

[1]. National Science Board (US). (2005). *National Science Board 2020 Vision for the National Science Foundation*. National Science Foundation.

[2]. National Science Board (US). (2007). *Enhancing support of Transformative Research at the National Science Foundation*. National Science Foundation.

[3]. Guo, Z., & Yang, A. (2019). Characteristics and experiences of the National Science Foundation of the United States. *Studies in Dialectics of Nature, 12*, 63-68.

[4]. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

[5]. Wang, W., & Ma, J. (2017). Topic detection and evolution analysis of research project based on LDA-study of projects on ocean acidification supported by NSF. *Journal of Intelligence, 36*(7), 34-39.

[6]. Trivedi, S. K., Patra, P., Singh, A., Deka, P., & Srivastava, P. R. (2022). Analyzing the research trends of COVID-19 using topic modeling approach. *Journal of Modelling in Management*, (ahead-of-print).

[7]. Tan, C., Chen, X., Liang, Y., & Li, Y. (2023). Sentiment analysis of onlookers on social media in privacy disclosure. *Information Science, 3*, 8-18.

[8]. Lee, K., Jung, H., & Song, M. (2016). Subject–method topic network analysis in communication studies. *Scientometrics, 109*, 1761-1787.

[9]. Yan, E. (2015). Research dynamics, impact, and dissemination: A topic-level analysis. *Journal of the Association for Information Science and Technology, 66*(11), 2357-2372.

[10]. Cretchley, J., Rooney, D., & Gallois, C. (2010). Mapping a 40-year history with Leximancer: Themes and concepts in the Journal of Cross-Cultural Psychology. *Journal of Cross-Cultural Psychology, 41*(3), 318-328.

[11]. Zhong, S., Geng, Y., Liu, W., Gao, C., & Chen, W. (2016). A bibliometric review on natural resource accounting during 1995–2014. *Journal of Cleaner Production, 139*, 122-132.

[12]. Huang, J., Peng, M., Wang, H., Cao, J., Gao, W., & Zhang, X. (2017). A probabilistic method for emerging topic tracking in microblog stream. *World Wide Web*, *20*, 325-350.

[13]. Zhang, S., Gao, Q., Ma, X., Wei, Y., & Yang, H. (2021). Study of learners' conversational behavior modeling in SPOC forums. *e-Education Research 42(11)*, 63-68.

[14]. Sivanandham, S., Sathish Kumar, A., Pradeep, R., & Sridhar, R. (2021). Analyzing research trends using topic modelling and trend prediction. *Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020*, 157-166.

[15]. Zhu, Q., Wan, J., Zhong, Y., Guan, Q., Zhang, L., & Li, D. (2020). Topic model for remote sensing data: A comprehensive review. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*,1373-1376. IEEE.

[16]. Yin, B., & Yuan, C. H. (2022). Detecting latent topics and trends in blended learning using LDA topic modeling. *Education and Information Technologies*, *27*, 12689–12712.

[17]. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, *32*, 288-296.

[18]. Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, 399-408.

[19]. Shin, J. S., Chatterjee, D., Krampner, J., & Virgin, V. (2020). *New York City Government Poverty Measure 2018: An Annual Report from the Office of the Mayor.* Retrieved from: https://www.nyc.gov/assets/opportunity/pdf/20_poverty_measure_report.pdf [accessed: 09 February 2023].