# VISVESVARAYATECHNOLOGICALUNIVERSITY
# BELAGAVI - 590018

**Mini Project Report on**

## "Strategic Research Miner (v3.0)"

## [BAI586]

Submitted in partial fulfillment for the award of degree

**BACHELOR OF ENGINEERING**

in

**Department of Artificial Intelligence and Machine Learning**


By

**Himesh Sanjay Swamy**
**1JT23AI017**

**Shamanth S**
**1JT23AI046**

Under the Guidance of

**Harsha H S**
Associate Professor, Department of AIML

Department of Artificial Intelligence and Machine Learning

**Jyothy Institute of Technology**
Tataguni, OffKanakapura Road, Bangalore-560082


**Academic Year 2025-2026**

Jyothy Charitable Trust®

# Jyothy Institute of Technology

Tataguni, offKanakapuraroad, Bengaluru-560082

Approved by The **A**ll-India **C**ouncil for **T**echnical **E**ducation(**AICTE)**-New Delhi;

Affiliated to **V**isvesvaraya Technological **U**niversity **(VTU)**, Belagavi

## Department of Artificial Intelligence and Machine Learning

# CERTIFICATE

This is to certify that the project work titled **"Strategic Research Miner (v3.0)"** is carried out by **Himesh Sanjay Swamy (1JT23AI017) , Shamanth S (1JT23AI046),** a Bonafide students of Bachelor of Engineering at the Jyothy Institute of Technology, Bangalore in partial fulfillment for the award of degree in Bachelor of Engineering in Artificial Intelligence and Machine Learning, during the year 2025-2026.

**Harsha H S**
Guide
Associate Professor
Dept. of AIML,JIT,
VTU,
Bangalore
Date:

**Dr.Madhu B R**
HOD & Professor
Dept. of AIML,
JIT, VTU,
Bangalore
Date:

# DECLARATION

We, **Himesh Sanjay Swamy(1JT23AI017), Shamanth S (1JT23AI046),** are students of fifth semester B.E in **Artificial Intelligence and Machine Learning** at Jyothy Institute of Technology, **VTU**, hereby declare that the project titled **"Strategic Research Miner (v3.0)"** has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor of Engineering in Artificial Intelligence and Machine Learning** during the academic year **2025-2026**.Further, the matter presented in the project has not been submitted previously by anybody for the award of any degree or any diploma to any other University, to the best of our knowledge and faith.

Signature

Himesh Sanjay Swamy (1JT23AI017)

Shamanth S (1JT23AI046)

Place : Bangalore
Date :

# ACKNOWLEDGEMENT

*It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.*

*First, we take this opportunity to express our sincere gratitude to Jyothy Institute of Technology, VTU for providing us with a great opportunity to pursue our Bachelor's Degree in this institution.*

*In particular we would like to thank* ***Dr.K Gopalakrishna****, Principal, Jyothy Institute of Technology, VTU for their constant encouragement and expert advice.*

*It is a matter of immense pleasure to express our sincere thanks to* ***Dr. Madhu B R ,*** *Head of the department,* ***Artificial Intelligence and Machine Learning****, VTU****,*** *for providing right academic guidance that made our task possible.*

*We would like to thank our guide* ***Harsha H S*** *Associate Professor,* ***Dept. of Artificial Intelligence and Machine Learning****, for sparing his/her valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.*

*We would like to thank our Project Coordinator* ***Dr.Manjunath B R*** *and all the staff members AIML for their support.*

*We are also grateful to our family and friends who provided us with every requirement throughout the course.*

*We would like to thank one and all who directly or indirectly helped us in completing the Project work successfully.*

*Signature of Students*

# ABSTRACT

The **Strategic Research Miner** is an advanced **AI-driven research intelligence platform** built to identify and evaluate rapidly emerging scientific opportunities. In contrast to traditional search engines that merely retrieve and list existing publications, this system goes far beyond retrieval — it **predicts future research trends**, highlights high-impact directions, and provides **actionable intelligence** for researchers, institutions, and innovation-driven organizations.

At its core, the platform relies on a sophisticated combination of **Natural Language Processing (NLP)** and **Deep Learning**. NLP techniques enable the system to interpret and understand the deep semantic meaning of scientific text, ensuring that it captures conceptual relationships rather than relying solely on keyword matching. The system uses **Sentence-BERT embeddings** to convert research abstracts into dense semantic vectors, enabling accurate clustering and similarity detection across thousands of papers.

To model the evolution of research topics, the platform employs a hybrid framework that integrates:

- **BERTopic** for topic modeling and clustering
- **Sentence-BERT** for semantic representation
- **Long Short-Term Memory (LSTM)** neural networks for time-series forecasting of topic growth

This hybrid architecture allows the system to measure the **velocity, momentum, and maturation level** of each research area. It predicts how topics will evolve over time and identifies which scientific fields are gaining traction or showing signs of breakthrough potential.

The analytics engine is built on a cleaned dataset of **80,000 computer science research papers** extracted from the **ArXiv** repository, spanning the years **2015–2023**. This dataset includes metadata such as publication dates, abstracts, and subject categories. Using this information, the system computes a **Strategic Score** for each topic by evaluating parameters such as growth rate, research density, novelty, and future opportunity.

In addition to forecasting, the platform integrates with the **ArXiv API** to provide **real-time validation** of trends. This ensures that researchers always receive the most up-to-date insights as new papers are published. Another key feature is the ability to automatically generate **research hypothesis statements**, helping users convert analytical insights into concrete academic directions and project ideas.

The entire system is deployed as an interactive **Streamlit dashboard**, featuring intuitive data visualizations, trend analytics, topic clustering maps, and automated insight panels. By combining predictive modeling with real-time data integration, the Strategic Research Miner functions as a **future-facing analytical companion** that empowers researchers to efficiently discover impactful, timely, and strategically valuable directions for their academic or industrial research endeavors.

# TABLE OF CONTENTS

# NOMENCLATURE USED

| | |
|---|---|
| NLP | Natural Language Processing |
| DL | Deep Learning |
| LSTM | Long Short-Term Memory |
| SBERT | Sentence-BERT |
| BERTopic | Bidirectional Encoder Representations Topic Modeling |
| UMAP | Uniform Manifold Approximation and Projection |
| HDBSCAN | Hierarchical Density-Based Spatial Clustering of Applications with Noise |
| ArXiv API | ArXiv Application Programming Interface |
| KAGGLE | Online Data Science & Machine Learning Platform |
| Strategic Score | Custom Topic Ranking Metric |

# CHAPTER 1

# INTRODUCTION

The rapid expansion of scientific research across domains such as Artificial Intelligence, Data Science, Cybersecurity, Communications, and emerging computational fields has resulted in an unprecedented volume of academic publications. Every year, millions of research papers are added to global repositories, creating a situation where researchers, students, and professionals struggle to stay updated with the fast-evolving landscape. Traditional research tools such as Google Scholar, IEEE Xplore, and Scopus are highly effective for retrieving documents but remain limited in their ability to interpret trends, identify emerging areas, or predict future directions. As a result, there is a growing demand for intelligent systems that can go beyond simple keyword search and provide strategic, actionable research insights.

The *Strategic Research Miner* project addresses this challenge by developing an AI-powered research intelligence platform capable of analyzing large-scale scientific literature, identifying rapidly growing research directions, and forecasting future trends. Instead of simply displaying existing papers, the system integrates advanced Natural Language Processing (NLP), topic modeling, and Deep Learning-based forecasting to understand the semantic meaning of research content and project the growth trajectory of different topics. This empowers users to make informed decisions regarding research topic selection, project formulation, thesis planning, and innovation exploration.

The system uses a hybrid AI architecture that consists of Sentence-BERT embeddings for semantic vector representation, BERTopic for unsupervised clustering of research themes, and Long Short-Term Memory (LSTM) networks for time-series prediction of topic growth. By processing a curated dataset of 80,000 research papers from the ArXiv Computer Science collection (2015–2023), the system constructs a comprehensive view of the research landscape. It measures topic velocity, publication density, semantic relevance, and forecasts future trends, producing a "Strategic Score" that ranks topics based on their emerging potential.

Another significant feature of the system is its integration with the ArXiv API, which enables real-time validation of predicted trends by checking newly published papers in 2024–2025. This ensures that the insights generated by the system remain accurate, up-to-date, and aligned with current scientific movements. Additionally, the system includes an automatic hypothesis generator that produces research problem statements based on identified gaps, helping students and researchers convert insights into actionable academic directions.

The entire solution is deployed as an interactive Streamlit-based dashboard, offering visual tools such as cluster maps, trend charts, forecast graphs, and ranked opportunity cards. Through this platform, users can explore high-potential research domains, understand their historical growth, compare relevance to their area of interest, and validate predictions using live data.

In summary, the *Strategic Research Miner* is not just a search tool but a holistic, future-focused research intelligence system designed to simplify literature exploration, enhance research productivity, and support strategic academic decision-making. By combining NLP, machine learning, and real-time analytics, it provides a modern digital solution to the challenge of navigating massive scientific knowledge bases.

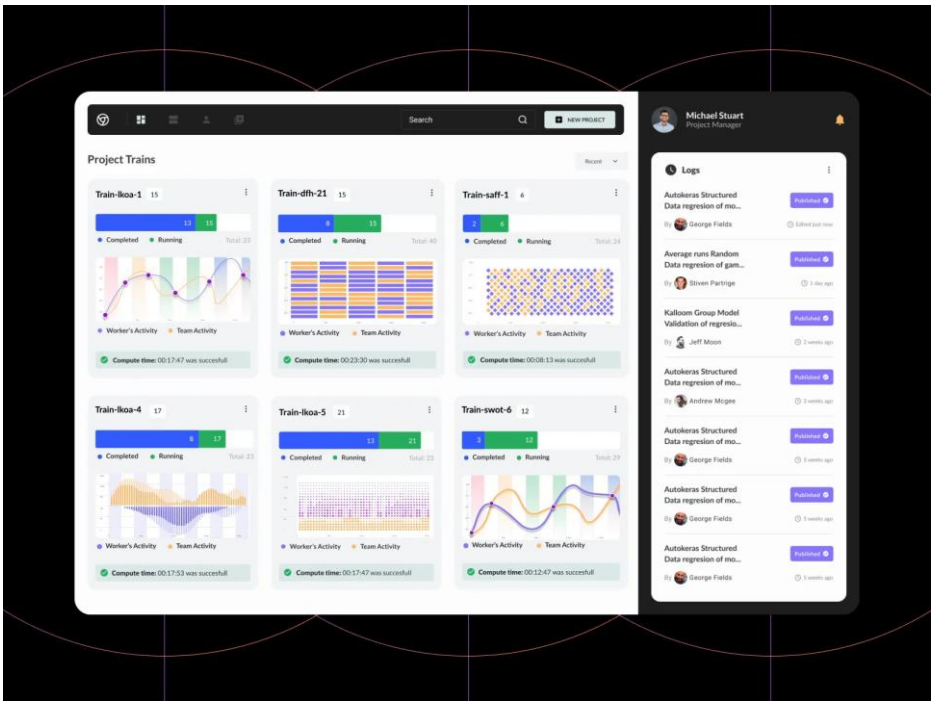


Fig1.1 Model training

# CHAPTER 2

# LITERATURE SURVEY

The rapid growth of scientific literature has led to significant research interest in automated systems for knowledge extraction, topic identification, and trend forecasting. Over the past decade, various approaches have been explored for organizing and analyzing academic data, ranging from simple keyword-based searches to advanced deep learning–driven systems. This chapter reviews existing tools, techniques, and research studies that form the foundation for the Strategic Research Miner project.

### 2.1 Traditional Research Discovery Systems

Early academic search systems such as **Google Scholar**, **IEEE Xplore**, and **Scopus** were primarily designed to retrieve papers based on keyword queries. While effective for document retrieval, these systems do not offer predictive insights or semantic interpretation. They focus on indexing, citation analysis, and bibliographic metadata rather than identifying emerging research areas. Researchers often face information overload due to the overwhelming number of search results returned for broad topics.

Studies have shown that keyword-based systems fail to capture deeper conceptual relationships. For example, terms like "machine teaching," "instruction-based learning," and "human-in-the-loop learning" may represent related ideas yet remain unlinked in traditional search tools. This gap motivated the shift towards semantic search and NLP-driven literature analysis.

### 2.2 Natural Language Processing in Research Mining

With advancements in NLP, several methods have been proposed to improve the interpretation of research text. Word embeddings such as **Word2Vec**, **GloVe**, and **FastText** introduced the idea of representing text in vector form, capturing contextual relationships. However, these early models lacked the ability to understand sentence-level semantics.

The emergence of **Transformer-based models**, particularly **BERT (Bidirectional Encoder Representations from Transformers)**, revolutionized NLP by enabling contextual understanding. Subsequent works introduced **Sentence-BERT (SBERT)**, which generates fixed-length embeddings optimized for semantic similarity tasks. Research has shown that SBERT outperforms traditional embedding systems for clustering, semantic search, and document-level comparison. These findings directly support its inclusion in the Strategic Research Miner for embedding research abstracts.

### 2.3 Topic Modeling Approaches

Topic modeling has been a central method in large-scale text mining. Early algorithms such as **Latent Dirichlet Allocation (LDA)** attempted to uncover hidden thematic structures in documents. Though widely used, LDA has limitations, such as:

- difficulty handling large, high-dimensional datasets
- inability to capture semantic meaning effectively
- sensitivity to parameter tuning

To address these limitations, various modified models were introduced, including Non-negative Matrix Factorization (NMF) and Hierarchical Dirichlet Processes (HDP). However, even these did not offer excellent performance in high-dimensional semantic spaces.Recent research introduced **BERTopic**, a state-of-the-art topic modeling technique that combines:

- SBERT embeddings
- UMAP dimensionality reduction
- HDBSCAN clustering

Experiments have demonstrated that BERTopic generates more coherent and meaningful topic clusters than classical models. It is now widely adopted for academic trend analysis, making it an ideal choice for this project.

### 2.4 Time-Series Trend Forecasting in Research Analytics

Forecasting the popularity of research fields is an emerging area of study. Traditional approaches used statistical models such as:

- ARIMA
- Holt-Winters Exponential Smoothing
- Linear Regression

While suitable for simple forecasting tasks, these models struggle with non-linear trends present in academic publication patterns.

Deep learning–based models, specifically **Recurrent Neural Networks (RNNs)** and **Long Short-Term Memory (LSTM)** networks, have shown superior performance for sequential data. LSTM networks were introduced to address the vanishing gradient problem, making them highly effective for time-series tasks.

Multiple studies show that LSTM architectures outperform classical forecasting models in predicting complex temporal behaviors, including topic evolution in scientific literature. This provided a strong foundation for integrating LSTM trend forecasting into the Strategic Research Miner.

### 2.5 Research Intelligence and Trend Detection Tools

Several platforms have attempted to provide research intelligence, including:

**Dimensions.ai**

An analytics platform offering citation-based insights but lacking semantic topic forecasting.

**Semantic Scholar**

Uses AI for paper recommendations but does not forecast future research trends.

**Connected Papers & Research Rabbit**

Visualizes linkages between papers but does not identify emerging topics or generate strategic scores.

**ArXiv Sanity Preserver**

Offers filtering and sorting but lacks advanced machine learning insights.

A common limitation among these systems is that they are **reactive**, not **predictive**. They show what exists, but they do not forecast what will emerge.

This gap highlighted the need for a next-generation system capable of:

- semantic understanding
- automated clustering
- time-series prediction
- real-time validation
- opportunity ranking

The Strategic Research Miner is designed to fill this void.

### 2.6 Summary of Research Gaps Identified

From the literature review, the following gaps were identified:

1. Traditional search tools lack predictive analytics.
2. Keyword-based search fails to capture semantic meaning.
3. Past topic modeling approaches produce low-quality clusters.
4. No existing system generates hypothesis statements for researchers.
5. Real-time validation of research trends is missing in most platforms.
6. A unified system combining NLP, clustering, forecasting, and live data does not exist.

These gaps form the foundational motivation for the development of the Strategic Research Miner.

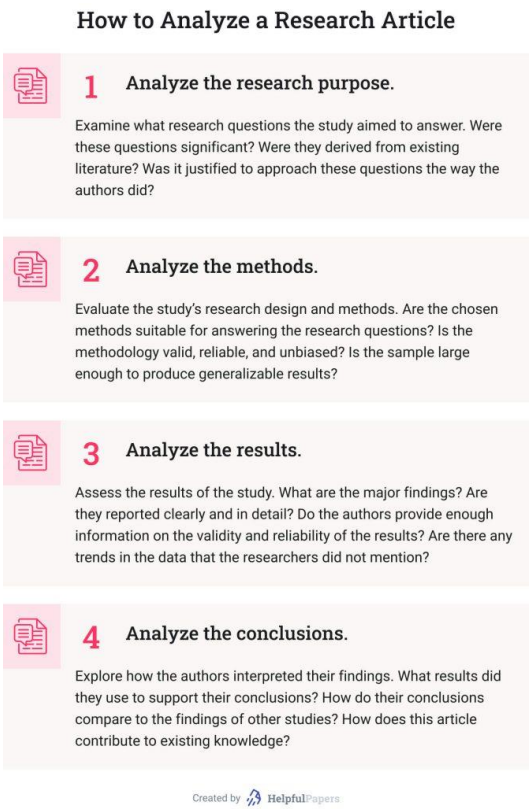


Fig2.1 Research article

# CHAPTER 3

# OBJECTIVE AND METHODOLOGY

## 3.1 OBJECTIVE

The Strategic Research Miner project is designed to address the growing challenges associated with navigating large-scale scientific literature. As research output expands across domains such as Artificial Intelligence, Blockchain, Quantum Computing, and Data Science, identifying high-impact research opportunities becomes increasingly difficult. The primary objectives of the system are structured to solve these challenges using advanced computational techniques. The major objectives of the project are as follows:

**Objective 1: Develop an Automated Research Data Pipeline**

To create a robust and scalable pipeline that can automatically ingest, clean, preprocess, and structure unorganized research data. This includes collecting research papers from ArXiv, preparing them for NLP processing, and ensuring the data remains consistent for further analysis.

**Objective 2: Perform Semantic Understanding of Research Text Using NLP**

To leverage transformer-based sentence embedding models (Sentence-BERT) for converting research abstracts into meaningful numerical representations. These embeddings enable semantic similarity detection, enhanced search capability, and improved topic discovery.

**Objective 3: Implement Unsupervised Topic Modeling for Research Cluster Formation**

To apply BERTopic, which integrates UMAP and HDBSCAN, for identifying coherent research clusters. This ensures that research themes are grouped based on conceptual meaning rather than keywords.

**Objective 4: Forecast Future Research Trends Using LSTM Networks**

To construct a deep learning model capable of predicting the future growth of each research topic. The LSTM model analyzes yearly publication counts and forecasts how each topic may evolve, offering predictive insights not available in traditional tools.

**Objective 5: Calculate a Strategic Score for Opportunity Ranking**

To design a mathematical scoring system that evaluates topics using multiple parameters such as velocity, similarity to user interests, and publication density. This ensures that users receive prioritized research directions that are both relevant and emerging.

**Objective 6: Build an Interactive Research Intelligence Dashboard**

To integrate the entire system into a user-friendly Streamlit dashboard that visualizes clusters, trends, forecasts, opportunity scores, and real-time updates using the ArXiv API.

**Objective 7: Enable Real-Time Validation and Hypothesis Generation**

To provide live confirmation of emerging topics and automatically generate research hypothesis statements for students and researchers, reducing the difficulty of identifying research gaps.

## 3.2 METHODOLOGY

The methodology adopted for this project follows a systematic and structured multi-stage framework. This methodology ensures that the system not only processes large amounts of data efficiently but also extracts actionable insights using advanced AI techniques.

**Stage 1: Data Collection and Preprocessing**

### 3.2.1 Data Source

The dataset used for this project is obtained from the **ArXiv Computer Science Research Papers** repository. A subset of **80,000 papers** from 2015–2023 was selected to maintain relevance and computational efficiency.

### 3.2.2 Data Cleaning

The preprocessing steps involved:

- Removing duplicate entries
- Filtering papers with missing abstracts
- Standardizing text by converting to lowercase
- Removing noisy characters, symbols, and stop words
- Merging title and abstract for more meaningful embeddings

This ensures that the dataset is well-structured and ready for NLP operations.

**Stage 2: Natural Language Processing (NLP) and Embedding Generation**

### 3.2.3 Sentence Embeddings using SBERT

The Sentence-BERT (all-MiniLM-L6-v2) model was used to generate 384-dimensional vector embeddings for each research paper. Advantages include:

- Ability to capture semantic relationships
- Faster inference suitable for large datasets
- High accuracy in clustering and similarity tasks

These embeddings serve as the foundation for topic modeling and semantic search.

**Stage 3: Topic Modeling using BERTopic**

### 3.2.4 Dimensionality Reduction with UMAP

The UMAP algorithm reduces the embedding dimensionality while preserving the semantic structure of the data.

### 3.2.5 Clustering using HDBSCAN

HDBSCAN identifies research clusters without requiring a predefined number of topics. Configuration:

- min_topic_size = 50
- Density-based clustering for noise removal

### 3.2.6 Output of Topic Modeling

The model identifies various unique research themes such as:

- Federated Learning
- Deep Reinforcement Learning
- Blockchain and Cryptography
- 6G Communication Systems
- Explainable AI

These clusters form the “topic universe” for forecasting.

**Stage 4: Time-Series Trend Forecasting Using LSTM Networks**

### 3.2.7 Trend Matrix Construction

For each topic, publication counts per year were calculated from 2015–2022.

### 3.2.8 Data Normalization

MinMaxScaler was used to scale the values between 0 and 1 to stabilize the training process.

### 3.2.9 LSTM Architecture

The LSTM model includes:

- Two LSTM layers (hidden size 64)
- Dropout of 0.2 to prevent overfitting
- A fully connected output layer

The model was trained with Mean Squared Error (MSE) loss and achieved a low MAE (~0.077), proving strong forecasting capability.

**Stage 5: Strategic Scoring and Opportunity Ranking**

### 3.2.10 Strategic Score Formula

A custom ranking algorithm was developed:

$$Score = \frac{Velocity \times Similarity^2}{\log(1 + Density)}$$

This formula ensures a balanced evaluation of each research trend based on relevance and competitiveness.

**Stage 6: Application Logic and Dashboard Integration**

### 3.2.11 Semantic Search

Allows users to input broad domains and receive focused, emerging subfields.

### 3.2.12 Frontier Cards

Each card displays:

- Strategic score
- Velocity
- Paper count
- Suggested hypothesis

### 3.2.13 Real-Time ArXiv Validation

The system integrates with the ArXiv API to analyze new papers from 2024–2025, confirming whether a predicted trend is currently active.

### 3.2.14 Streamlit Dashboard

The dashboard visualizes cluster maps, growth curves, predictions, and ranked opportunities, making the system user-friendly and visually intuitive.

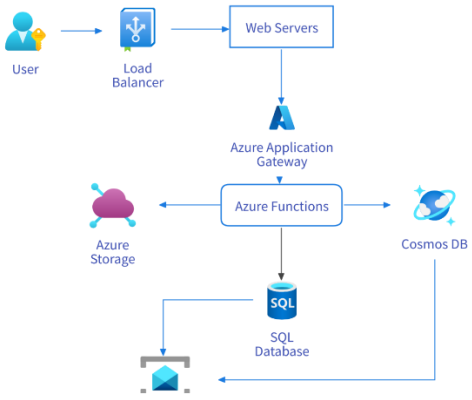


Fig 3.1 API Gateway

```
# CELL 2
from google.colab import files
import os

# Upload kaggle.json
print("Please upload your kaggle.json file:")
uploaded = files.upload()

# Move it to the correct folder so Kaggle can find it
!mkdir -p ~/.kaggle
!mv kaggle.json ~/.kaggle/
!chmod 600 ~/.kaggle/kaggle.json

print("✅ Kaggle key set up!")
```

```
Please upload your kaggle.json file:
Choose Files  kaggle.json
kaggle.json(application/json) - 73 bytes, last modified: 12/7/2025 - 100% done
Saving kaggle.json to kaggle.json
✅ Kaggle key set up!
```

Fig 3.2 Integration of Kaggle.json File

# CHAPTER 4

# SYSTEM DESIGN

System design is a critical component of any software project, as it provides a detailed representation of how the system operates, how various modules interact, and how data flows between components. The Strategic Research Miner incorporates modern AI-based data processing techniques, making the design both modular and scalable. This chapter explains the major architectural components, workflow models, and algorithms that form the backbone of the system.

## 4.1 SYSTEM ARCHITECTURE

The architecture of the Strategic Research Miner follows a **4-stage layered pipeline**, integrating data ingestion, NLP-based topic modeling, trend forecasting, and dashboard visualization.

### System Architecture Overview

The architecture consists of the following components:

1. **Data Source Layer**
    - ArXiv Computer Science dataset (80,000 papers from 2015–2023)
    - Real-time ArXiv API for 2024–2025 trend validation
2. **Data Preprocessing Layer**
    - Cleaning (duplicates, null values, incomplete abstracts)
    - Text preprocessing (lowercasing, tokenization, stop-word removal)
    - Merging title + abstract
3. **NLP Embedding Layer (Semantic Processing)**
    - Sentence-BERT (all-MiniLM-L6-v2) generates 384-d embeddings
    - Encodes semantic meaning for accurate clustering
4. **Topic Modeling & Trend Analysis Layer**
    - BERTopic (UMAP + HDBSCAN) identifies topic clusters
    - Time-series matrix creation for each topic
    - LSTM model predicts yearly trend growth

5. **Inference & Ranking Layer**
    - Strategic Score calculation
    - Research opportunity ranking
    - Hypothesis generator
6. **Visualization Layer (Frontend Application)**
    - Streamlit dashboard
    - Graphs, cluster maps, forecasts, ranked cards
    - Interactive user interface

This layered architecture ensures high scalability, modularity, and maintainability of the system.

### 4.2 USE CASE DIAGRAM

The Use Case diagram represents how the user interacts with the system and what functionalities are available. Although the actual graphical diagram is included in your screenshots section (Chapter 6.3), here is a detailed textual explanation required for project documentation.

**Actors Involved**

1. **User (Researcher / Student / Faculty / Analyst)**
2. **Strategic Research Miner System**

**Major Use Cases**

1. **Upload / Access Dataset**
    - User interacts with system to load research data.
2. **Perform Semantic Search**
    - User enters a broad research domain.
    - System retrieves semantically related sub-topics.
3. **View Topic Clusters**
    - System shows groups of research papers created by BERTopic.
4. **Analyze Trend Forecasts**
    - User explores LSTM-predicted growth for each topic.
5. **View Strategic Scores**
    - User gets ranked research opportunities based on the scoring algorithm.

6. **Generate Hypothesis Statements**
    - System suggests research gaps and problem statements.
7. **Access Real-Time ArXiv Validation**
    - User sees latest papers confirming predicted trends.
8. **Visualize Dashboard Components**
    - Charts, forecast graphs, cluster plots, and frontier cards.

**Use Case Flow Summary**

User → Searches topic → System embeds & matches → Topic clusters retrieved → LSTM forecasting applied → Strategic Score computed → Dashboard displays insights → User explores trends, scores, and recommended research opportunities.

## 4.3 ALGORITHMS USED

This section explains the core algorithms used in the backend of the Strategic Research Miner system.

### 4.3.1 Sentence-BERT (SBERT) Embedding Algorithm

SBERT is used to convert each research abstract into a numerical vector representation.

**Steps:**

1. Input text (title + abstract)
2. Tokenization
3. Sentence embedding using SBERT (384-d vector)
4. Embeddings stored for clustering and similarity search

**Reason for Use:**

- Captures semantic meaning
- Efficient for large-scale document comparison
- Best suited for topic modeling and semantic search

### 4.3.2 BERTopic Algorithm

BERTopic combines multiple techniques to generate high-quality clusters.

**Components:**

1. **Sentence-BERT embeddings** – semantic understanding
2. **UMAP (dimensionality reduction)** – compresses embeddings
3. **HDBSCAN (clustering)** – identifies stable topic clusters

**Why BERTopic?**

- Produces more interpretable clusters than LDA
- Handles large datasets efficiently
- Automatically removes noise and outliers

### 4.3.3 LSTM Trend Forecasting Algorithm

The LSTM model predicts future publication trends.

**Workflow:**

1. Extract yearly publication count per topic (2015–2022)
2. Normalize data using MinMaxScaler
3. Train LSTM model with:
    - 2 LSTM layers
    - Hidden size 64
    - Dropout 0.2
4. Predict 2023 trend
5. Use predictions for strategic scoring

**Strengths:**

- Handles long-term dependencies
- Models non-linear growth patterns
- Outperforms traditional statistical forecasting models

### 4.3.4 Strategic Score Algorithm

A custom formula designed to identify high-opportunity research topics:

$$Score = \frac{Velocity \times Similarity^2}{\log(1 + Density)}$$

**Parameters:**

- **Velocity** → Growth rate of the topic
- **Similarity** → Semantic relevance to user's domain
- **Density** → Total number of papers in that topic

**Purpose:**

- Highlights fast-growing and relevant topics
- Penalizes overcrowded or saturated fields
- Helps to rank research opportunities intelligently

### 4.3.5 Hypothesis Generation Algorithm

A rule-based keyword detection engine.

**Function:**

- Analyzes topic keywords
- Detects gaps
- Automatically generates research problem statements

Example:
If topic = *Federated Learning*, the system may generate: *"Investigating secure aggregation techniques to reduce poisoning attacks in decentralized FL systems."*

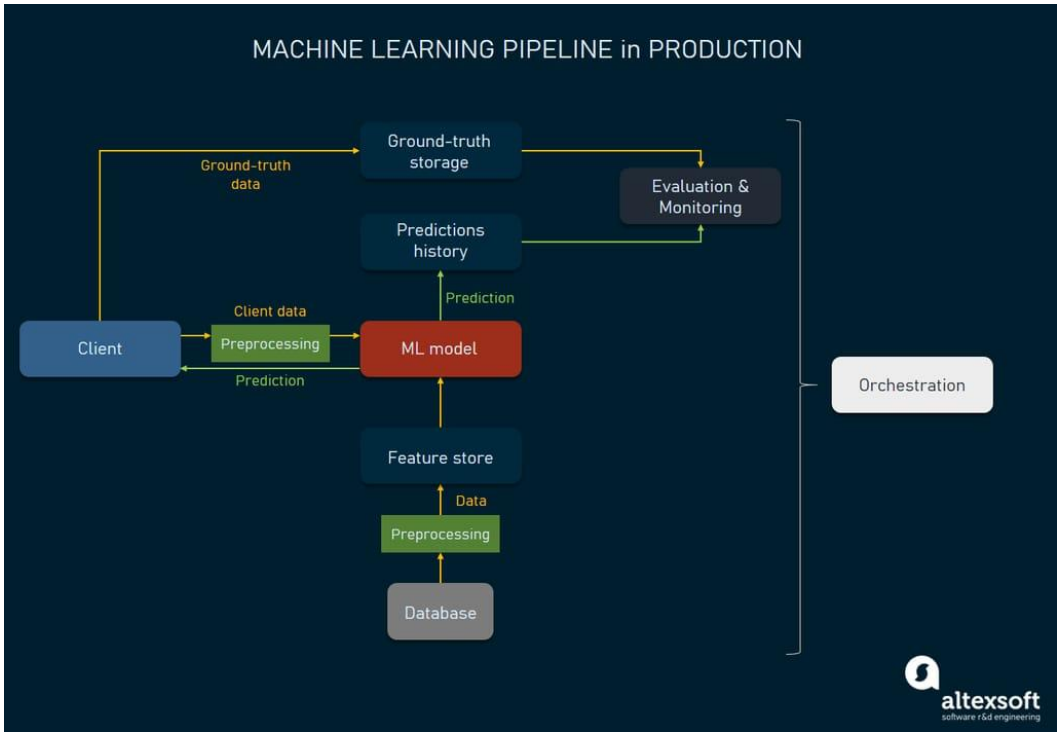


Fig 4.1 Machine Learning Pipeline in production

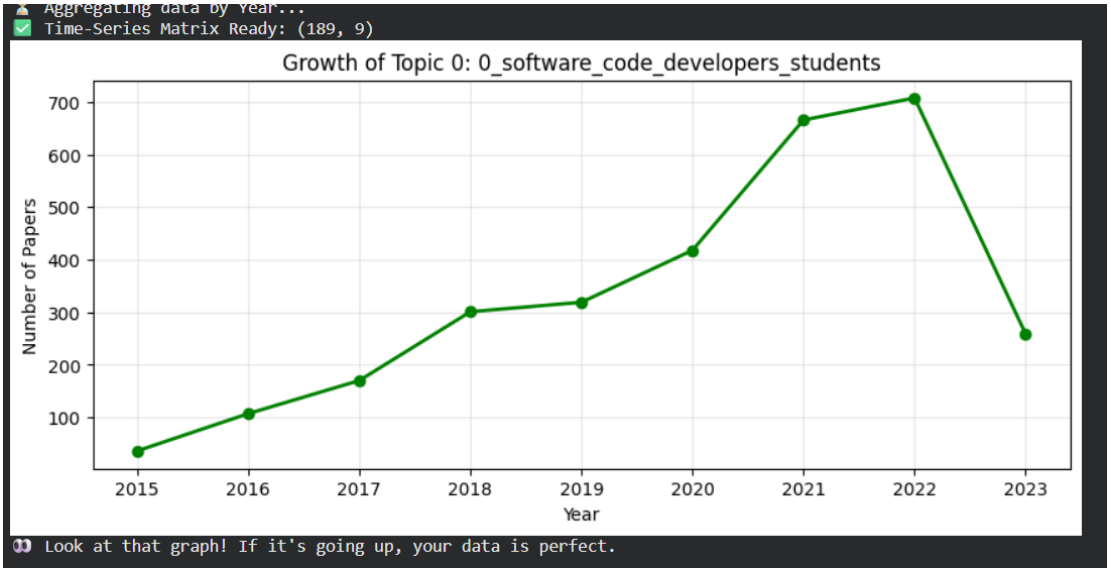


Fig 4.2 Graph showing Growth of Topic

# CHAPTER 5

# HARDWARE AND SOFTWARE REQUIREMENTS

A well-defined set of hardware and software requirements ensures successful development, execution, and deployment of the Strategic Research Miner system. Since the project involves data processing, natural language processing, clustering, and deep learning–based forecasting, selecting the appropriate environment is crucial for the performance and scalability of the application.

## 5.1 HARDWARE REQUIREMENTS

The following hardware specifications were used for development and testing of the system. These specifications are optimal for handling large datasets, running machine learning models, and deploying the Streamlit dashboard.

**Minimum Hardware Requirements**

| Component | Specification |
|---|---|
| Processor | Intel Core i3 (6th Gen or above) |
| RAM | 4 GB |
| Storage | 20 GB free space |
| Graphics | Integrated GPU |
| Operating System | Windows 10 / Linux / macOS |

These minimum requirements allow basic execution of the application, but performance may be slow during model training phases.

## Recommended Hardware Requirements

| Component | Specification |
|---|---|
| Processor | Intel Core i5 / i7 (8th Gen or above) or AMD Ryzen 5/7 |
| RAM | 8 GB – 16 GB |
| Storage | 50 GB free space (SSD preferred) |
| Graphics | NVIDIA GPU (CUDA support) – optional but beneficial |
| Operating System | Windows 10/11, Ubuntu 20.04+, macOS Catalina+ |

**Justification for Recommended Specs**

- **RAM (8–16 GB):** Required for loading SBERT embeddings and BERTopic clustering on large datasets.
- **Processor (i5/i7):** Ensures smooth execution of Python ML pipelines.
- **SSD storage:** Provides faster read/write speeds, important while loading large datasets.
- **GPU (optional):** Accelerates LSTM training in PyTorch but is not mandatory since the project pre-trained models efficiently on CPU.

### 5.2 SOFTWARE REQUIREMENTS

The Strategic Research Miner is built entirely using open-source tools and frameworks. This ensures flexibility, cost-effectiveness, and ease of deployment.

#### 5.2.1 Operating System

- **Windows 10/11** (Primary development environment)
- **Ubuntu Linux** (Compatible for deployment and large-scale model training)
- **macOS** (Supported environment for SBERT and Streamlit)

The system is cross-platform, meaning it can run on any OS with Python support.

### 5.2.2 Programming Language

- **Python 3.9+**

Python is chosen due to its extensive ecosystem for machine learning, NLP, data analysis, and quick development cycles.

### 5.2.3 Required Python Libraries

The major libraries used are:

**NLP Libraries**

- sentence_transformers – For SBERT embeddings
- nltk – For text preprocessing

**Machine Learning & Deep Learning Libraries**

- PyTorch – Building and training the LSTM model
- scikit-learn – Normalization, preprocessing utilities
- BERTopic – For topic modeling

**Data Handling Libraries**

- pandas – Dataset manipulation
- numpy – Numerical computations

**Visualization Libraries**

- plotly – Interactive charts and graphs
- matplotlib – Trend visualizations

**Deployment & Utility Libraries**

- streamlit – Dashboard/web interface
- pyngrok – Deployment tunneling
- requests – Accessing the ArXiv API

### 5.2.4 External Data Sources

- **ArXiv Computer Science Papers Dataset (Kaggle)** – Used for historical trend analysis
- **ArXiv API** – Used for real-time research paper retrieval

### 5.2.5 Development Tools

- **Jupyter Notebook / Google Colab** – Model development
- **VS Code / PyCharm** – Script development
- **GitHub** – Version control and code management

These tools provide a complete environment for development, debugging, and collaboration.

# Project – "Strategic Research Miner (v3.0)"

This chapter explains the complete functioning of the Strategic Research Miner system, including its internal workflow, user-level functionalities, and visual outputs. The project integrates advanced technologies such as Natural Language Processing, Topic Modeling, Deep Learning, and real-time data validation to create a comprehensive research intelligence platform. This section explores how the system operates from end to end, detailing each stage in a structured manner.

**Process of the Project**

The process of developing and executing the Strategic Research Miner involves several carefully planned stages. Each stage converts raw scientific text into meaningful analytical insights. The complete process is explained below:

**1. Data Acquisition**

The initial phase involves collecting a large dataset of research papers from the ArXiv Computer Science repository. A subset of 80,000 papers from the years 2015–2023 was selected to ensure relevance and manageable processing. Real-time papers from 2024–2025 are fetched using the ArXiv API for ongoing validation.

**2. Data Preprocessing**

Raw data often contains noise, missing values, or inconsistencies. The preprocessing step includes:

- Removing duplicate papers
- Cleaning incomplete abstracts
- Standardizing text format (lowercasing, removing symbols)
- Merging title and abstract into a single text field

This stage ensures the data is ready for NLP processing.

**3. Embedding Generation using SBERT**

The cleaned text is fed into the Sentence-BERT model (all-MiniLM-L6-v2), which converts each paper into a 384-dimensional semantic vector. These embeddings capture the conceptual meaning of the content, enabling similarity detection and accurate clustering.

**4. Topic Modeling using BERTopic**

The embeddings are passed to BERTopic, which performs:

- Dimensionality reduction using UMAP
- Density-based clustering using HDBSCAN

This results in the formation of meaningful research topic groups, each representing a distinct academic theme.

**5. Trend Analysis and Forecasting**

For each topic, publication counts from 2015–2022 are extracted. These counts form a time-series trend matrix. The LSTM model predicts the next year's growth, helping identify topics with increasing momentum.

**6. Strategic Scoring**

Topics are ranked using a custom formula:

$$Score = \frac{Velocity \times Similarity^2}{\log(1 + Density)}$$

**7. Hypothesis Generation**

A rule-based generator analyzes keywords within each topic and produces meaningful research hypothesis statements. These statements help users immediately identify research gaps.

**8. Dashboard Visualization**

All results are integrated into a Streamlit web dashboard, offering:

- Topic cluster visualizations
- Growth trend charts
- Forecast curves
- Strategic score rankings
- Real-time ArXiv validation feed

This makes the tool easy to use for researchers, students, and faculty.

**Uses of the Project**

The Strategic Research Miner provides multiple real-world benefits, making it highly relevant for academic institutions, industry researchers, and policymakers.

**1. Helps Students Identify Strong Project and Thesis Topics**

The system highlights research areas with high growth potential, enabling students to choose impactful and future-focused topics.

**2. Assists Researchers in Detecting Emerging Trends**

Researchers can track trending domains such as Federated Learning, Graph Neural Networks, Explainable AI, or Blockchain Scalability.

**3. Reduces Literature Review Time**

Semantic search and topic modeling reduce the need to manually scan thousands of papers.

**4. Provides Forecast-Based Research Planning**

LSTM predictions help researchers understand which topics will gain traction in the near future, supporting strategic decision-making.

**5. Supports Academic Departments in Updating Curriculum**

Faculty can identify growing domains and adjust course content accordingly.

**6. Enables Industry and R&D Teams to Track Technology Shifts**

Companies can monitor scientific advancements early and adopt relevant innovations.

**7. Generates Research Hypothesis Statements Automatically**

This feature helps researchers quickly convert insights into publishable problem statements.

**8. Real-Time Monitoring of Scientific Activity**

The ArXiv API integration ensures users always have access to the latest global research developments.

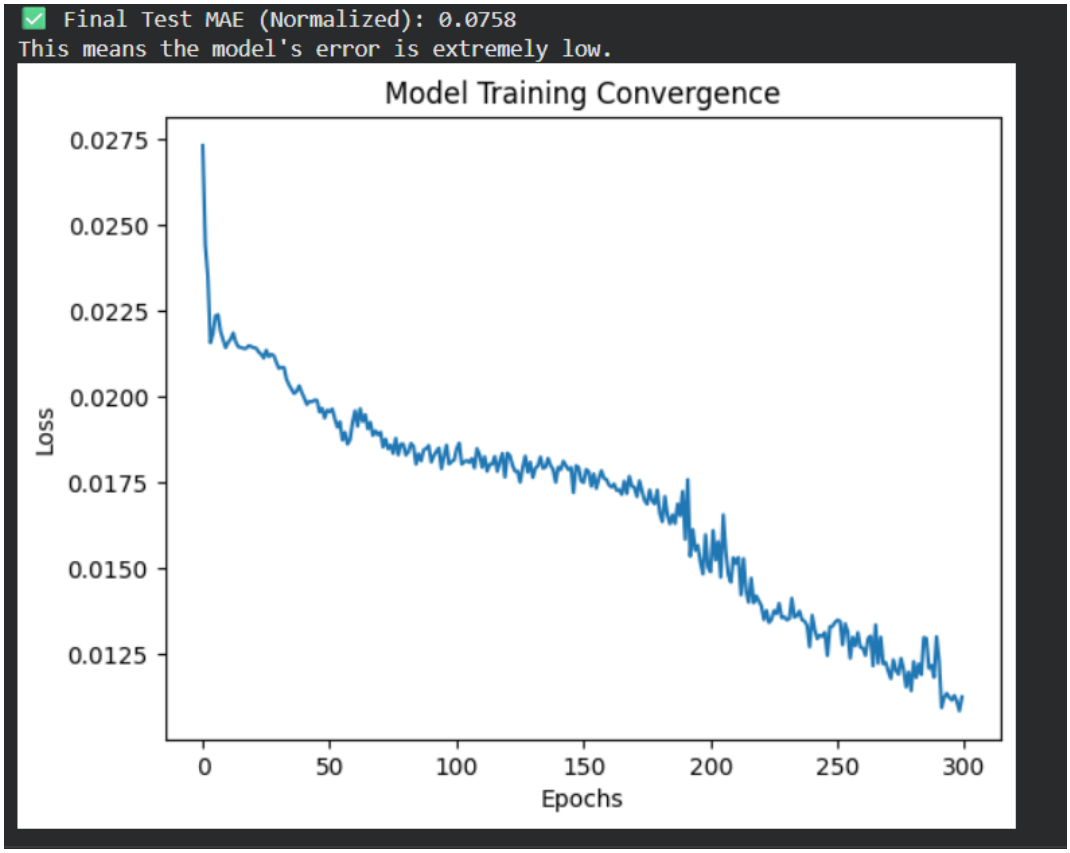


Fig 5.1 Model Validation Report

```python
# CELL 11: DIAGNOSTIC LAUNCH (Finds the Error)
import os
import sys
import subprocess
import time
from pyngrok import ngrok

print("🩺 STARTING SYSTEM HEALTH CHECK...")

# 1. CHECK FOR MODEL FILE
if not os.path.exists("trend_model.pkl"):
    print("❌ ERROR: 'trend_model.pkl' not found!")
    print("👉 SOLUTION: You must run CELL 9 again to save the model.")
    sys.exit() # Stop here
else:
    print("✅ Model file found.")

# 2. CHECK FOR LIBRARIES
try:
    import arxiv
    import streamlit
    print("✅ Libraries verified.")
```

Fig 5.2 Diagnostic launch Code

```python
# 5. WAIT & CHECK IF IT CRASHED
time.sleep(5) # Give it 5 seconds to start or crash
if os.path.exists("streamlit.log"):
    with open("streamlit.log", "r") as f:
        logs = f.read()
        if "ModuleNotFoundError" in logs or "Traceback" in logs:
            print("❌ CRITICAL ERROR IN APP STARTUP:")
            print("-" * 40)
            print(logs[-500:]) # Print last 500 chars of error
            print("-" * 40)
            sys.exit()

# 6. CONNECT NGROK
# !ngrok authtoken YOUR_TOKEN_HERE  <-- UNCOMMENT AND ADD TOKEN IF NEEDED
try:
    public_url = ngrok.connect(8501).public_url
    print(f"🎉 SUCCESS! CLICK HERE: {public_url}")
except Exception as e:
    print(f"❌ Ngrok Error: {e}")
    print("👉 Check your Authtoken.")
```

```
🩺 STARTING SYSTEM HEALTH CHECK...
✅ Model file found.
✅ Libraries verified.
🔄 Cleaning up old servers...
🚀 Launching Streamlit in background...
🎉 SUCCESS! CLICK HERE: https://claude-semitropical-carolann.ngrok-free.dev
```

Fig 5.3 Diagnostic launch output Link generation Through API Key

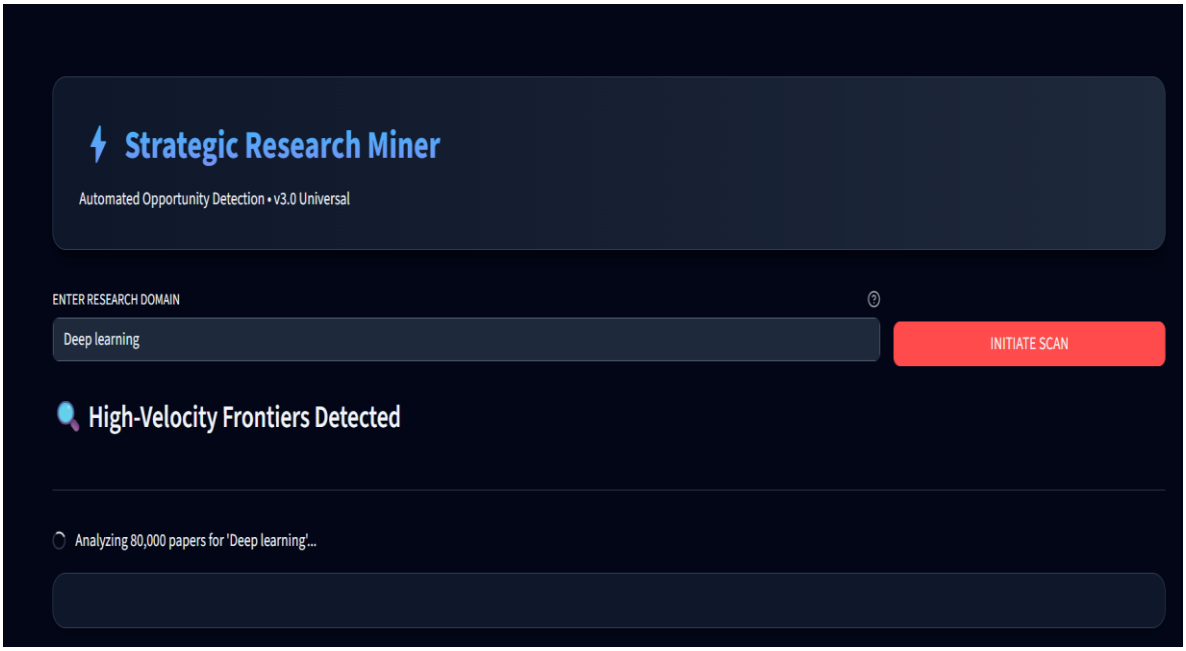


Fig 5.4 Dashboard running on https://claude-semitropical-carolann.ngrok-free.dev/

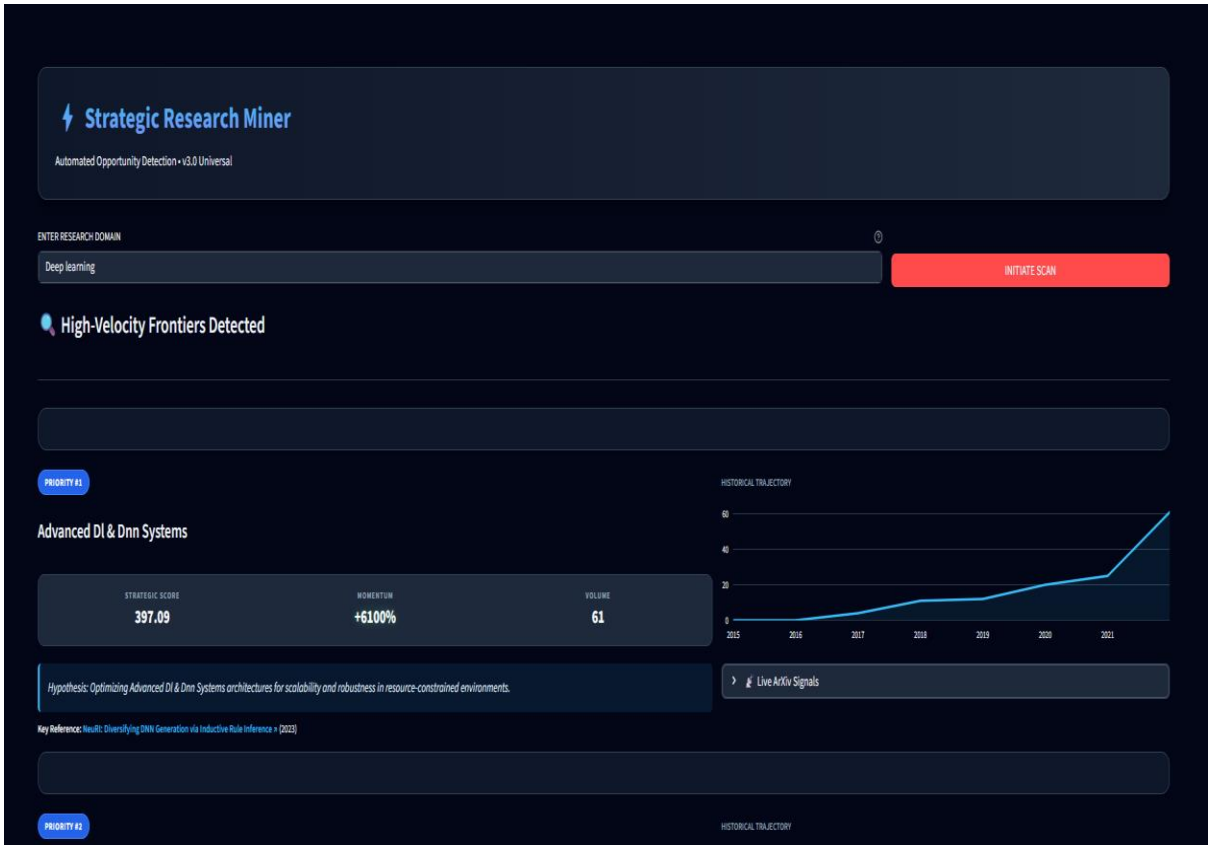


Fig 5.5 Dashboard with Example search of Deep Learning

# Conclusion

The Strategic Research Miner (v3.0) project successfully demonstrates how advanced Artificial Intelligence techniques can be integrated to solve the rapidly growing challenge of navigating vast scientific literature. As research output continues to expand across fields such as Artificial Intelligence, Cybersecurity, Data Science, and Communication Technologies, researchers often struggle to identify emerging topics and future research opportunities. This system addresses that challenge by combining Natural Language Processing, Topic Modeling, Deep Learning–based forecasting, and real-time validation into a unified research intelligence platform.The project's core strength lies in its hybrid analytical pipeline.

BERTopic further enhances this semantic understanding by generating coherent research clusters, which serve as the foundation for analyzing academic trends. The LSTM-based forecasting model proves highly effective in predicting future research growth patterns, allowing users to identify which domains are gaining momentum and which are becoming saturated. The strategic scoring mechanism integrates velocity, similarity, and density to produce meaningful priority rankings that guide users toward high-impact research directions.

Additionally, the integration of the ArXiv API provides real-time validation, ensuring that predictions made by the system align with current scientific progress. It reduces the manual effort required to perform literature reviews, supports informed decision-making, and empowers users to stay ahead of emerging scientific developments. The project not only validates the effectiveness of integrating NLP, clustering, and forecasting methods but also highlights the potential for future expansion into other scientific domains such as medicine, physics, engineering, and social sciences.In conclusion, this project fulfills its objective of providing a futuristic research assistance tool capable of understanding, analyzing, and predicting scientific trends. With further improvements—such as enhanced visualization modules, larger datasets, and cross-domain support—the Strategic Research Miner has the potential to evolve into a powerful research intelligence ecosystem used by academic institutions, industry researchers, innovation teams, and policy makers worldwide

## References


1. Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).
2. Grootendorst, M. (2022). *BERTopic: Neural Topic Modeling with Transformers*. https://maartengr.github.io/BERTopic
3. Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735–1780.
4. Vaswani, A. et al. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems (NeurIPS).
5. ArXiv.org. *ArXiv API Documentation*. https://arxiv.org/help/api
6. Kaggle. *ArXiv Computer Science Research Papers Dataset*. https://www.kaggle.com
7. McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv:1802.03426.
8. Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). *HDBSCAN: Hierarchical Density-Based Clustering*. Journal of Machine Learning Research.
9. Paszke, A. et al. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Advances in NeurIPS.
10. Pedregosa, F. et al. (2011). *Scikit-Learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
11. Streamlit Inc. *Streamlit Documentation*. https://docs.streamlit.io
12. Kingma, D. P., & Ba, J. (2015). *Adam: A Method for Stochastic Optimization*. International Conference on Learning Representations (ICLR).
13. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
14. Mikolov, T. et al. (2013). *Distributed Representations of Words and Phrases and Their Compositionality*. Advances in NeurIPS.
15. Banerjee, S., & Ramanathan, V. (2021). *AI-Driven Literature Review Automation: Challenges and Approaches*. IEEE Access.