

# SmartSearchify: AI-Enabled Semantic Search for eSankhyiki Portal

Rahul Khichar  
IIT Gandhinagar  
Gandhinagar, Gujarat, India  
23110264@iitgn.ac.in

Vadithya Harsha Vardhan  
Nayak  
IIT Gandhinagar  
Gandhinagar, Gujarat, India  
23110349@iitgn.ac.in

Sailaja Mudu  
IIT Gandhinagar  
Gandhinagar, Gujarat, India  
23110288@iitgn.ac.in

## ABSTRACT

The eSankhyiki Portal provides access to India's official statistics, but its dropdown-based search makes finding relevant data difficult. This project proposes an AI-powered semantic search system that understands natural language queries and maps them to the correct statistical indicators and filters. By using Natural Language Processing (NLP) and machine learning, the system will improve accessibility, usability, and the overall user experience of the portal.

## KEYWORDS

Semantic Search, Natural Language Processing, eSankhyiki Portal, AI-based Query Mapping

## 1 INTRODUCTION

The eSankhyiki Portal provides structured access to India's official statistics. However, users must manually navigate dropdown menus, making it difficult for non-experts to find relevant data.

This project aims to develop an AI-based semantic search system that allows users to enter queries in natural language and get accurate results with direct links to relevant database indicators and filters. This will improve usability, enhance data discoverability, and make accessing information easier.

## 2 RELATED WORK

- **Existing AI Models:** Advanced models like BERT and GPT have improved semantic search capabilities.
- **Improved User Experience:** Previous AI-based search systems have enhanced accessibility and engagement.
- **Limited Use in Statistics:** Applying AI-based semantic search to structured statistical databases is still an emerging area.
- **Building on Prior Work:** This project extends existing research to improve data retrieval in statistical databases.

## 3 METHODOLOGY

- **Collecting Data:** Gather and organize statistical indicators from the eSankhyiki Portal using web scraping or an API.
- **Understanding Queries:** Use Natural Language Processing (NLP) to interpret user queries and match them with the correct indicators.
- **Finding Similar Data:** Compare user queries with existing data using similarity measures like cosine similarity and word embeddings.
- **Improving Search Results:** Suggest relevant queries and rank results based on accuracy to enhance user experience.

- **Training and Testing:** Train and evaluate the model to ensure it performs better than traditional keyword-based searches.

## 4 DATASETS AND LIBRARIES

- **Datasets:** eSankhyiki database, open-source NLP datasets.
- **Libraries:**
  - Web Scraping: beautiful-soup, selenium
  - NLP Processing: NLTK, SpaCy, Hugging Face Transformers.
  - Similarity Search: FAISS.
  - Model Training: TensorFlow/PyTorch.
  - Backend: FastAPI/streamlit
  - Frontend: HTML, CSS, js

## 5 TIMELINE

- **Week 1:** Literature review, data collection, exploratory analysis.
- **Week 2:** Model selection, baseline NLP implementation.
- **Week 3:** Model training and fine-tuning.
- **Final week:** Testing, evaluation, optimization.

## REFERENCES

- [1] eSankhyiki Portal. [Online]. Available: <https://sankhyiki.in/>