AI-enabled Semantic Search for National Classification of Occupation (NCO)

1. Track

Data Collection and processing

2. Description

The National Classification of Occupation (NCO) is a standardized system for classifying occupations in India, aligned with the International Standard Classification of Occupations (ISCO). The current version, NCO-2015, includes detailed descriptions of 3,600 civilian occupations across 52 sectors, structured through an 8-digit hierarchical code. Presently, the only method available for users (e.g., survey enumerators, employment officers) to identify relevant occupation codes is via keyword-based search in static documents (PDFs). This approach:

- Requires exact keyword match

- Offers no semantic understanding of queries
- Demands extensive familiarity with the occupation taxonomy

- Is time-consuming, error-prone, and not scalable

For example, to assign the occupation code for a "Sewing Machine Operator," a user must know its classification hierarchy across divisions and sub-groups a challenging task without automation. This complexity can result in incorrect classification, leading to data quality issues that hinder policy planning and productivity. If unresolved, this challenge will continue to affect the efficiency and accuracy of official statistics, directly impacting national data systems and evidence-based governance.

3. Expected Outcomes/Solutions

Participants are expected to develop a prototype application that:

- Ingests and indexes NCO-2015 data
- Accepts free-text input from users (e.g., job title, description)
- Returns top N matching occupation codes with:
- Semantic relevance ranking
- Confidence scores

- Supports error handling and fallback suggestions
- Integration with voice input or multilingual support
- Admin panel to update or revise the occupation database
- Audit trails for search history or manual override

4. Relevance to National Priorities or Ongoing MoSPI Initiatives

This use case aligns with MoSPI's commitment to leveraging AI and frontier technologies for improving the quality, speed, and usability of official statistics. By enabling accurate, semantic, and intelligent occupation code assignment, it contributes to:

- High-quality, real-time data collection
- Efficient resource deployment in large-scale surveys
- Reduced manual errors and training overhead
- Modernization of classification systems

5. Background Resources or Datasets

MoSPI will provide:
- NCO-2015 data (PDF/Excel format)
  - Volume I: Mapping with NCO-2004
  - Volume II: Detailed job descriptions
- Sample queries and mapped codes (if available)

Participants may use pre-trained language models (e.g., BERT) and standard NLP resources, but final applications must operate within the context of Indian occupational taxonomies.

6. Key Features Required

Data Ingestion & Processing
- Convert NCO datasets into structured formats (e.g., JSON, CSV)
- Normalize text for uniformity
- Preserve and represent hierarchy (4-digit groupings to 8-digit codes)

Semantic Search Implementation
- Generate embeddings using NLP models (e.g., BERT, Sentence Transformers)
- Store indexed embeddings in a fast-retrieval system (e.g., FAISS)
- Return results with ranked relevance and confidence scores

Synonym & Context Handling
- Map synonyms and variations (e.g., "tailor" vs. "sewing machine operator")

- Create a synonym/related term bank

Query Interface
- Accept text/voice input
- Display top matches and allow user selection
- Include fallback and error messages

- Support for multilingual input (Hindi and regional languages)
- Dashboards for usage statistics, performance, and audit trail
- API-based integration with MoSPI survey apps or portals

7. **Impact Potential**

The solution has high impact potential by:

- Reducing manual effort for enumerators and increasing accuracy
- Improving data consistency across regions and surveys
- Accelerating survey preparation and reducing training time
- Enabling scalable, intelligent systems for national classification tasks

It directly strengthens MoSPI's capacity to deliver timely, high-quality, and policy-relevant official statistics.

Improved Efficiency: An AI/ML solution will reduce the cost and manual effort required for classifying the scanner data, if used in future.

Enhanced Accuracy: With automated classification and utilization of scanner data will more accurately capture the inflation or deflation which is critical for different policies in the country.