Education 4.0: Leveraging AI for intelligent course selection

A paper on suggested approaches for an implementation done between Nagarro and Indian Institute of Technology (Madras) to empower students to choose the right courses and improve their career roadmap with AI intervention.

nagarro

nagarro

Table of Contents

Executive summary: AI is revolutionizing the way we learn 2

Understanding the problem 2

- The challenges 3
- The Ground Reality and Datasets 3

Approaches 4

- Deep learning Approach 5
- Explicit Semantic Analysis (ESA) Approach 6

Results 7

Conclusion 8

1. Executive summary

Artificial intelligence is revolutionizing education with optimized career roadmaps.

Artificial intelligence is bringing interventions in the education sector across many different areas like learning, teaching, assessment/evaluation, and institutional operations. These interventions are improving the overall experience for students as well as teachers. One of the areas where universities face challenges, is the course and subject selections during any program.

Optimal course selection can help a student have better learning and hence an improved career roadmap. This paper will walk you through the analysis and research conducted between Nagarro experts and the Indian Institute of Technology Madras (IIT Madras) and the suggested approaches required for an implementation that aims to empower students to choose the best-suited courses with the help of AI intervention. The key is to understand the similarities between different courses and then decide on the course selection.

2. Understanding the problem

IIT Madras offers several interdisciplinary courses for undergraduate and postgraduate students in its curriculum. It allows the students to choose courses from various disciplines for their academic year. New courses are added by every department throughout the year that are open to all students. The number of courses available per year is generally very high to allow manual analysis or comparison for course content similarity or overlap. For instance, in one of the 2019 semesters, around 500 courses were offered to the students.

So, there arises a need to analyze and compare courses effectively. The system can be automated and scaled as required for many courses. This mechanism could help in several use cases, such as:

- Assist students from different departments in selecting a balanced set of course curriculum.
- Help faculty to assess course similarity/overlap for a student.
- Assist faculty members who are willing to float a new course in identifying overlapping courses.

2.1 The challenges

The following challenges cropped up while evaluating the course content for similarities.

- **Analyzing a huge volume of content**
  Courses generally contain a large volume of text content such as course descriptions, course material, references, etc. Analyzing such a course's content is challenging because these texts are often complex sentences and contain domain-specific information. Analyzing and comparing these courses on a scale is another challenge, as the accuracy of such analysis is highly dependent on the quality of the data.

- **Analyzing and comparing courses from different domains**

  While collectively analyzing courses from various departments, there can be many subjects/domains the content can fall under. Often the words or phrases used in one domain may hold a completely different meaning in another domain. The importance of a given word or phrase could depend on the context of the domain in question.

  For instance, the word "regression" might have a higher weightage than "data" in the "Pattern Recognition and Machine Learning" context. Domain information is always better understood by its terminology, synonyms, and abbreviations that are popular in that domain. For instance, *features* and *attributes* are used interchangeably in the CSE domain, which may not be the case in other domains. Also, sometimes some concepts and ideas are understood, implied, and not always mentioned in course content.

This makes it very important that the analysis approach used isn't term specific; rather the approach considers the context of the sentence being used.

2.2 Ground reality and datasets

A course is often described by using multiple fields, such as its title, the course content, the set of skills a student is expected to acquire by course completion, a list of reference books, etc. These details are often presented on the course page of the course.

For example, AS2040 - Flight Dynamics course details are shown in Figure 1.

Course Number	AS2040
Course Name	Flight Dynamics I
Department Code	AE
Faculty Name	JOEL GEORGE M
Description	Introduce students to concepts and problem solving in aircraft performance and stability
Content	Forces and moments acting on a vehicle in flight. Translational equations of motion of a grid flight vehicle. International standard atmosphere. Various types of drags. Drag polar vehicles from low speeds to hypersonic speeds. Review of the variation of thrust / power and SFC with altitude and velocity, for various air breathing engines. Performance of airplane in level flight, glide, climb, accelerated flight, turn, take-off and landing. Flight limitations and envelopes. Flight-testing: Altitude definitions, Speed definitions, Air speed, altitude, and temperature measurements; Flight determination of drag polar. Basic concepts of trim and stability; Aircraft static stability; Neutral and maneuver points and their determination in-flight.
Reference Books	-

Figure 1: AS2040 - Flight Dynamics course details

To analyze the data, the IIT team provided an extract of 1043 courses and their details. An extract must be pulled from the course details in their academic system. Every semester, the courses often are changed/updated, new courses are added, and old ones are removed. Nagarro was mostly dependent on the curated course data provided by IIT.

Nagarro and IIT teams have worked on different approaches for the proof-of-concept (POC). Depending on the approach, the teams have considered suitable methods of pre-processing the data, as detailed in the next section of the paper.

3. Approaches

3.1 Deep Learning Approach

Deep learning models are generally trained using large sets of examples and neural network architectures with many layers. These models achieved state-of-the-art effectiveness for many problems, including those in Natural Language Processing.

Compared to other methods, the main advantage of using a deep learning approach is its ability to analyze unstructured text data better. This approach is more suitable for understanding long domain-specific data, as mentioned in the Challenges section. With more data, the performance of such models can be improved proportionally. The ability to analyze text by generating features automatically through data examples makes it trainable and scalable easily over time.

Two Iterations

Nagarro adopted the deep learning approach, and during this exercise, we defined the scope and executed the POC in two iterations to generate quick results for the business to evaluate.

1. The **first iteration** of test results was generated using a pre-trained model called Universal Sentence Encoder. Also, two dynamic reports were designed to view the results. The first report listed the top five similar courses for a selected course. The second report was to compare the similarity of the two selected courses.

   Google's Universal Sentence Encoder (USE) is a model architecture that encodes sentences into high-dimensional vectors. These vectors can be used in multiple tasks such as sentiment analysis, sentence clustering, text classification, semantic analysis, and other NLP tasks. Instead of returning word embeddings, USE works on encoding sentences that offer better performance and applicability. Many versions of pretrained USE are available, and it is the most downloaded model available at Tensorflow Hub. Cosine similarity was used to measure the similarity between courses.

2. The **second iteration** of test results was generated using a Bidirectional Encoder Representations from Transformers (BERT) model finetuned on labeled IIT courses for Computer Science and Engineering (CSE) and Electric Engineering (EE).

   BERT has recently introduced improved state-of-art benchmarks for many natural language processing tasks. The model is designed to pre-train deep bidirectional representations from unlabeled text, which uses its transformer architecture to use left and right directional context in all layers. The pre-trained model also allows finetuning by adding additional layers to allow usage for a wide range of text analysis tasks without making many changes to the architecture.

Data Preparation

Data preparation for the initial modeling was performed with the help of the IIT team as follows:

- Data cleanup was performed to remove duplicates and incomplete course details. This action resulted in a set of 520 unique courses. 'Course Name,' 'Course Description,' and 'Course Content' columns were used for initial modeling.
- Labeled data, *G*, of similar courses in Computer Science Engineering, CSE, and Electrical Engineering, EE, the department was collated by the IIT team. Preparing and reviewing the labeled data was an extensive process. So only courses from the CSE and EE departments were selected. This data was used as the ground truth for initial modeling for finetuning the BERT model.

$G = \{(x, y) \mid y \text{ overlaps with } x \text{ and } x \in \{CSE, EE\}\}$

Reports

The report lets the user select a course and lists the top 5 similar courses from the analysis. The score represents the similarity; it is scaled between 0-1, where a score of 0 denotes dissimilar and 1 denotes more similar (see Figure 2).

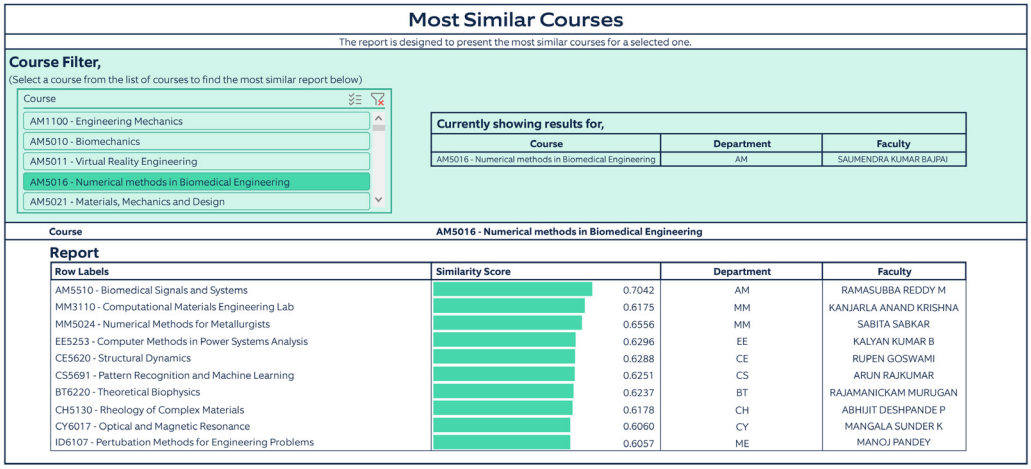

Figure 2: Course Similarity Score

Compare course similarities

The report lets the user select two courses and the similarity comparison is shown to the user, along with its score. In Figure 2, the user can see a field-wise comparison that helps understand the similarities.

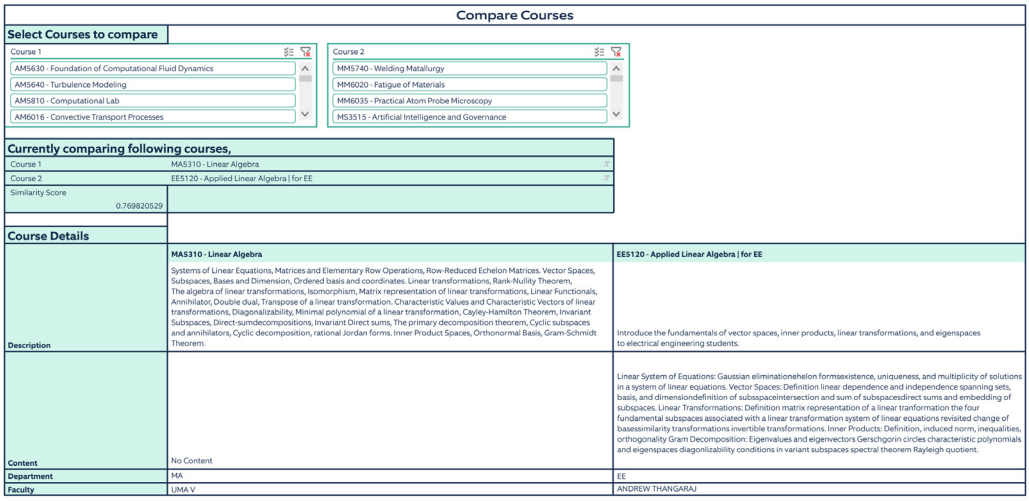

Figure 3: Compare course similarities

3.2 Explicit Semantic Analysis (ESA) Approach

ESA is a non-introspective approach used to represent documents in a space spanned by orthogonal Wikipedia concepts. Expressing words and documents of a corpus in terms of Wikipedia concepts enables us to incorporate world knowledge from Wikipedia. The IIT team now pursued ESA approach.

For this, a wiki TF-IDF matrix is constructed. It indicates the TF-IDF weight of every word from the Wikipedia vocabulary in every Wikipedia article. Hence, the i^{th} row of this matrix can be interpreted as a vector representation of the i^{th} word from the Wikipedia vocabulary in terms of the Wikipedia articles (referred to as the Wikipedia concepts). Such word vectors are called concept vectors.

Finally, the vector representation of a document in the corpus of concern is computed by taking a linear combination of the concept vectors of the constituent words with weights as the TF-IDF score of the words in that corpus. ESA has been shown to tackle the problem of synonymy well since it can capture world knowledge from Wikipedia.

- **wiki-ESA**

  To build a subset of Wikipedia for world knowledge in ESA, Wikipedia was queried using the Wikipedia python library with queries as the title of the courses offered by the institute. This method has the advatage that it makes ESA computationally feasible since it is not required to handle data from the entire Wikipedia and captures only domain-specific Wikipedia articles (equivalently, domain-specific knowledge). Taking the example of a query course q, similar courses are arranged in decreasing order of their cosine similarity with q.

- **OUCC-ESA**

  The team attempted to capture more domain-specific knowledge from Other University Courses' Contents (OUCC). For this, Google was queried with the titles of the courses offered at IIT Madras and they scraped the content available on the website that belonged to the following universities: Massachusetts Institute of Technology (MIT), Stanford University, New York University (NYU), IIT Kanpur, IIT Delhi, IISc Bangalore, IIT Kharagpur, IIT Roorkee, IIT Bombay. Each webpage extracted using the mentioned method corresponds to a concept in the ESA procedure.

Phrase extraction

Phrase extraction is another approach where the set of phrases in a corpus is defined as, {(x, y) | Mutual-Information (x, y) > θ}. θ is a hyperparameter that acts as a threshold on the mutual information values.

Given such pairs of words, any occurrence of a word pair 'x y' (where x is immediately followed by y) is replaced by x_y. For instance, if
Mutual-Information (machine, learning) > θ, all occurrences of 'machine learning' are replaced by 'machine_learning' and are therefore treated as a single token in the corpus instead of two. In other words, this word pair will now be treated as a phrase instead of two separate words, and hence will not coalesce with the constituent words' vector representations. Python also provides an internal implementation for extracting phrases of arbitrary length in the Phraser module.

As part of this activity, the team wanted to arrive at a list of pre-existing courses offered at IIT Madras that have a significant overlap with a course that the institute or a faculty member wants to introduce. The available courses were expected to have the following fields:

1. The course title
2. A course description terms of the topics covered
3. A list of objectives to be met by the students post course completion
4. A list of reference books

For the sake of experimentation, the team adapted the leave-one-out strategy wherein they considered every existing course as a query with the corpus of courses as all the pre-existing courses, but the query. The Information Retrieval (IR) system then retrieves courses similar to every query course and averages the performance over all such queries. The ordered list generated by the IR system is to retrieve the most relevant courses at the top ranks for every course query.

Results

Deep learning results were generated based on the labeled data generated by the IIT team after iteration 1. The labeling was done by the IIT faculty considering two factors.

1. If the two course contents being compared have any overlap.
2. If one of the courses is a pre-requisite in the curriculum for the other course being compared.

Based on the results from iteration 1 and considering the factors, in iteration 2, a new model is prepared by fine-tuning the BERT model. The fine-tuning was done on a sample training set and generated results on a sample test set of the labeled data.

Approach	Methods	Accuracy Scores
Deep learning	Iteration 1 - USE	65%
Deep learning	Iteration 2 - BERT - Fine-tuned with given labeled data	80%

Approach	Methods	nDCG Scores
Explicit Semantic Analysis	Wiki ESA	83%
Explicit Semantic Analysis	OUCC ESA	82%

Conclusion

Nagarro is working towards many different AI interventions in the education domain. Nagarro's contribution to this project has helped apply AI capabilities in the education domain. These kinds of problems are relevant for many educational institutions and similar solutions can help them improve student experience significantly.

For IIT Madras, this collaboration has helped explore various approaches available for course analysis. In the long term, this would help the institute to build a course analysis solution that can optimize course design and course plan, while helping students with course selection. A prototype of this system has been deployed and is being used by the stakeholders of the institute.

About the authors

Nagarro:

Surya Kiran

Ramesh Soni

Sachin Virmani

Sagar Papneja

IIT M:

Adwait Parsodkar

Shanu Kumar

Special acknowledgement

Shania Mitra

Tapish Garg

Shashank Patil

Special thanks to **Prof Anil Prabhakar**, **Prof Sutanu Chakraborti**, **Monika Gupta** and **Manohar Venugopal** for guidance and coordination.

About Nagarro

Nagarro is a global digital engineering leader with a full-service offering, including digital product engineering, digital commerce, customer experience, AI and ML-based solutions, Cloud, immersive technologies, IoT solutions, and consulting on next-generation ERP. We help our clients become innovative, digital-first companies through our entrepreneurial and agile mindset, and we deliver on our promise of thinking breakthroughs.

We have a broad and long-standing international customer base, primarily in Europe and North America. This includes many global blue-chip companies, leading independent software vendors (ISVs), other market and industry leaders, and public sector clients.

Today, we are over 19,500+ experts across 35 countries, forming a Nation of Nagarrians, ready to help our customers succeed.

For more information, visit www.nagarro.com