Retaining Disadvantaged Students Using a BERT-based Recommender System

Muhammad Ashfakur Arju, John Sheppard, Md. Asaduzzaman Noor, Carina Beck, Durward Sobek
Montana State University
Bozeman, MT, USA
{first.last}@montana.edu

***Abstract*—Recent initiatives in United States university systems have been focusing on the problems associated with first-year students dropping out in high numbers. At Montana State University, a pilot program is underway to develop strategies for improving undergraduate student retention and reducing the time to graduation. For the pilot program, students found to be socioeconomically and academically disadvantaged are targeted to participate in the strategies designed to mitigate their disadvantaged state. Through the university's "Persistence to Degree" initiative, these strategies have included forming student cohorts that take first-year classes together, thus promoting a sense of purpose and belonging. This paper presents one such strategy whereby the cohorts are formed based on mutual interests, and common first-year "core" courses (i.e., general education courses) are recommended for the members of the cohorts to take together. The approach involves employing a BERT-based topic model to generate a social network, from which communities are extracted. These communities are constructed based on student interests, expressed through a set of "minute essays," and a hybrid content-based/collaborative filtering method is employed to pair courses taken by past students with similar interests. Given that this project is still in the pilot stage, this paper focuses on the methodology and underlying ethical issues, rather than specific results since collecting results on the method's effectiveness would require more of a longitudinal design. Even so, initial results show promise in the proposed methodology.**

***Index Terms*—BERTopic, Social Network Analysis, Community Detection, Recommender Systems, Purpose and Belonging**

I. Introduction

Universities within the United States are struggling to improve student retention and time-to-degree [1]. One of the issues facing new students is developing a sense of purpose and belonging in a new environment and obtaining support as they navigate this environment. One approach to addressing concerns in purpose and belonging and student support is in recommending appropriate courses to incoming students. With the increasing number of course options and the diverse backgrounds of students, personalized course recommendation systems and the formation of communities among these students are expected to help address such concerns. In this paper, we present an approach that is focused on improved community engagement and student welfare. In particular, we focus on incoming students facing either academic, engagement, or financial challenges in the hope of helping them integrate and succeed in the university environment. For our approach, we utilize data from the Hilleman Scholars program at Montana State University, where selected students are asked to write short minute essays completed as part of their first-year experience. These responses provide a valuable dataset for understanding student interests, goals, and challenges that can then be analyzed to improve purpose, belonging, and student support.

Previous work on course recommendation systems have used a variety of methods to provide personalized suggestions to students, including collaborative filtering, content-based filtering, and hybrid approaches. While these methods have been effective in recommending courses based on historical data such as course ratings, there has been limited focus on leveraging unstructured textual data for building personalized recommendations. Our objectives are two-fold. First, we seek to provide a means for students to be grouped into communities of mutual interest to improve a sense of belonging in a new environment. Second, we seek to use this information on common interests to build a data-driven framework that helps match students in these communities to relevant courses.

To achieve our objectives, we employ a multi-step process. First, we utilize the BERTopic model [2] to extract meaningful topics from the textual responses of students admitted to the university's Hilleman Scholars program. This topic modeling process [3] enables us to identify themes and patterns that represent the unique characteristics of different students. Next, we construct a social network based on these topic models, where nodes represent students and edges represent topic similarities. To identify communities in the network, we apply the hierarchical Louvain algorithm [4].

When an incoming student joins the Hilleman Scholars program and provides their responses to the brief essay questions, we match them to an appropriate community within the social network by finding the most similar student in the network. This is achieved by computing the similarity between the new response and existing nodes using cosine similarity and selecting the node with the highest similarity. Once the community is identified, we query the grade database of the members within the community to retrieve their academic performance in various courses. By multiplying similarity scores with grade point averages, we recommend N number of core courses to new students.

Contributions: Our first contribution is a targeted approach to enhancing retention and belonging among academically and socioeconomically disadvantaged students. By focusing on this underserved group, our work addresses a critical gap in university retention efforts through personalized course recommendations. Our second contribution is methodological: we propose a hybrid BERTopic-based recommender system that combines unstructured student responses with academic performance. By extracting student interests via transformer-based topic modeling, building a semantic simi-

larity network, and applying hierarchical Louvain clustering, we generate personalized course recommendations.

II. Related Work

There exists several methods for making course recommendations. Many methods rely on temporal registration and grading data while others rely on course ratings, student feedback, and other relevant scores [5]–[8].

Both methods described in [6] and [7] use a combination of two-tier collaborative filtering and recommendation. Elbadrawy and Karypis, use matrix factorization, collaborative filtering, and popularity to recommend top-n courses [6]. They group students and courses based on their features, but their method becomes less effective if any of those groups are absent in the dataset. Professor ratings can also be used as a quality filter to exclude courses taught by poorly rated professors [7], but this method does not account for how much bias is introduced from student ratings.

Zhang *et al.* demonstrated that collaborative filtering can use past data to find patterns among students, though it struggles with problems like cold start and sparsity [9]. Thorat *et al.* suggested that hybrid methods, which combine collaborative and content-based filtering, can help overcome these issues [10]. Such methods use multiple data sources to make recommendations more reliable.

Another popular approach is topic modeling, which draws its inspiration from Probabilistic Latent Semantic Analysis (PLSA) [11]. Earlier topic modeling methods such as Latent Dirichlet Allocation (LDA) were based on latent variable discovery [12]. This is similar to the approach of PLSA, which analyzed the occurrence of words in unstructured textual documents. Building on this idea, LDA introduced a generative process to explain how documents and words might be generated in a probabilistic framework. As LDA yields a generative model it provides flexibility of tuning, interpretability, and scalability, but when working with a small corpus of data, LDA falls short as it heavily relies on word frequency to learn the latent topics. Subsequently, transformer-based models, such as BERTopic [2], were developed. BERTopic, which is the method used in this paper, generates contextual word embeddings that capture semantic relationships between words, even when the dataset is small.

Jiang *et al.* used LDA to analyze text responses and identify student interests [13]. This approach was demonstrated to work well for unstructured data. Recent studies then tried to improve personalized recommendations by adding time and context to the analysis [14].

Other machine learning techniques have also been used in course recommendations. Ng and Linn showed that adding sentiment analysis and survey results can make recommendations more accurate. Systems based on matrix factorization and deep learning have created highly personalized suggestions [15]. Support Vector Machines (SVM) is utilized to group students for personalized recommendations [16]; however, scalability remains a challenge for these methods.

Proceeding from approaches that employ social network analysis, graph-based approaches are used. These methods aim to find relationships between textual representations. For example, Noor *et al.* also focused on constructing social networks entirely based on topic similarity [17]. By using Jensen-Shannon divergence and community detection algorithms like the Louvain algorithm [18], [19], this method highlighted interdisciplinary collaboration opportunities.

Some studies have proposed new ways to recommend courses for specific scenarios. For example, CrsRecs uses sentiment analysis and survey data to create detailed student profiles [15]. Similarly, Jiang *et al.* developed an LDA-based system to recommend online courses based on user interest models [13]. These systems demonstrate how specialized methods can address unique challenges in course recommendation.

In this work, we consider the problem of making course recommendations from a perspective of enhancing social good. In particular, our goal is to make recommendations that promote a sense of purpose and belonging among the students while also matching students to courses of interest that improve the likelihood of success. To that end, we propose a new approach to course recommendations where we combine topic modeling with social network analysis to group students based on text responses to mini-essay and interest survey questions. We then use prior student performance data to improve the accuracy of our recommendations. This approach introduces a fresh way to match students with suitable courses, focusing on unstructured text data and community-based grouping.

III. Problem Statement

Within the larger context of seeking to promote purpose, belonging, and student success, we aim to solve the technical problem of identifying communities from freshman student responses to essay and survey questions. These communities are based on similarities of topics contained in their text responses. Our end-goal is to recommend courses that meet their specific interests and needs but also promote success. Our basic approach is to build these communities by analyzing unstructured textual data rather than relying on predefined relationships or explicit survey scores. This approach ensures that recommendations are based on shared themes and patterns in the students' responses.

We hypothesize that a social network constructed using topic modeling and similarity-based measures will identify student communities effectively. These communities can then be used to recommend courses that align with the academic potential and interests of incoming students. By incorporating grade performance data from similar students, we aim to enhance the relevance of these recommendations. This framework presents an approach to use unstructured text data for creating personalized academic advising for incoming, disadvantaged students. We note, however, that testing this hypothesis requires tracking students over the duration of their time at the university. As such, this paper focuses more on the methodology being developed for a pilot study than on the specific results obtained.

IV. Datasets

Throughout the process of developing our framework, we used data collected from the Hilleman Scholars program program at Montana State University. The first dataset (denoted ESSAYS) consists of questionnaire responses from students who participated in the program between 2016 and 2024.

These data consist of 649 entries, representing responses to 25 questions, including both open-ended and structured formats. The responses enabled gaining insights about students' academic interests, career goals, personal motivations, and voluntary work performed.

The second dataset (denoted GRADES) includes grade information for the same students. These data consist of 14,945 entries, providing detailed records for each student of the courses taken, final course grades, cumulative GPAs, credit hours, math and writing placement scores, and standardized test scores.

Together, the two datasets offer a comprehensive view of student responses and academic performance. The ESSAYS dataset provides unstructured textual data suitable for topic modeling, while the GRADES dataset serves as a quantitative basis for validating course recommendations. By combining the datasets, we analyzed how students' self-reported goals and motivations aligned with their academic performance, enabling us to develop a robust recommendation system. The resulting system is designed to guide incoming students toward courses that best match their aspirations and potential, promote academic community engagement, and enable successful progress towards degree completion.

V. METHODOLOGY

Our system collects questionnaire responses on an application given to incoming freshmen as part of the Hilleman Scholars program that focuses on helping socio-economically disadvantaged students. The questionnaire contains 25 questions aimed at understanding the student's academic interests, career goals, and motivations. Each response is tied to a unique student identifier, which appears as a node identifier in the derived social network. After fine-tuning the BERTopic model, we process the responses to extract topics which are based on similarity and word embedding in the documents. These topics are used to build a social network, where each student is represented as a node, and edges are formed based on topic similarities using cosine similarity.

To identify groups within the network, we apply the hierarchical Louvain algorithm [18], which groups students into communities and sub-communities. The questionnaire data is also linked to the GRADES dataset, which is used to validate and refine the recommendations by analyzing trends in course performance within the identified sub-communities.

A. BERTopic Model

BERTopic [2] is a newer topic modeling algorithm introduced in 2020 that leverages transformer-based embeddings, dimensionality reduction, and clustering to discover latent topics in textual data. Unlike traditional topic modeling approaches such as LDA, BERTopic uses semantic embeddings generated by transformers based on BERT [20]. This allows BERTopic to model text with contextual meanings effectively, which is particularly useful for unstructured data like short student responses.

Formally, given a collection of textual documents $D = \{d_1, d_2, \ldots, d_n\}$, where d_i represents a single document (e.g., a student response), the first step in BERTopic is to transform the text into high-dimensional embeddings. Using a pre-trained transformer model T, each document d_i is mapped to a vector $e_i = T(d_i),\ e_i \in \mathbb{R}^d$, where d is the embedding dimensionality.

TABLE I: Topic and word representation from BERTopic

Topic	Count	Representation
-1	193	['work', 'help', 'school', 'make', 'life', 'would', 'time', 'family', 'get', 'go']
0	140	['life', 'help', 'work', 'make', 'get', 'school', 'go', 'people', 'time', 'would']
1	113	['work', 'help', 'school', 'time', 'life', 'year', 'family', 'make', 'go', 'get']
2	44	['work', 'Hilleman', 'hard', 'life', 'help', 'family', 'Montana', 'world', 'make', 'story']
3	34	['work', 'help', 'school', 'make', 'life', 'time', 'take', 'people', 'go', 'hard']
4	26	['want', 'work', 'go', 'get', 'help', 'school', 'make', 'time', 'like', 'know']
5	13	['work', 'help', 'Montana', 'make', 'goal', 'time', 'want', 'teach', 'community', 'like']
6	13	['help', 'people', 'life', 'want', 'work', 'make', 'time', 'go', 'year', 'like']
7	12	['work', 'school', 'time', 'Montana', 'make', 'get', 'college', 'family', 'help', 'year']
8	11	['work', 'get', 'help', 'class', 'would', 'school', 'make', 'year', 'go', 'time']

Since clustering in high-dimensional space can be computationally expensive, BERTopic employs Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction [21]. The embeddings e_i are mapped to a lower-dimensional space $z_i = \text{UMAP}(e_i),\ z_i \in \mathbb{R}^m,\ m < d$. UMAP retains the global and local structures of the original embeddings while making the data more amenable to clustering.

After dimensionality reduction, BERTopic applies Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [22] to identify groups of documents that form coherent topics. The algorithm clusters the reduced embeddings z_i, assigning a topic label t_i to each document d_i: $t_i = \text{HDBSCAN}(z_i),\ t_i \in \{1, 2, \ldots, k\}$ where k is the total number of discovered topics. HDBSCAN is particularly well-suited for this task because it can handle noise in the data, effectively identifying meaningful clusters while ignoring outliers.

To provide a meaningful representation for each topic, BERTopic computes a class-based Term Frequency–Inverse Document Frequency (c-TF-IDF) score. This score quantifies the importance of a term w within a topic T:

$$\text{c-TF-IDF}(w, T) = \frac{\text{TF}(w, T)}{\sum_{w' \in T} \text{TF}(w', T)} \log \frac{|D|}{|\{d \in D : w \in d\}|},$$

where $\text{TF}(w, T)$ is the term frequency of w in topic T, $|D|$ is the total number of documents, and $|\{d \in D : w \in d\}|$ is the number of documents containing the term w. The terms with the highest c-TF-IDF scores for a topic are selected as its representative keywords.

Once the topics are identified and represented, BERTopic assigns each document d_i a probability distribution $\theta_{i,j}$ over the topics T_j. The final topic assignment for a document is determined by selecting the topic with the highest probability:

$$\hat{T}(d_i) = \underset{j}{\text{argmax}}\ \theta_{i,j}$$

This ensures that each document is associated with the topic most relevant to its content.

We leveraged BERTopic model to extract latent topics that reflect student's academic interests, career goals, and personal motivations. Table I summarizes the output of the BERTopic model, which extracted latent topics from student responses. Each topic is represented by a unique identifier, the number of documents assigned to it (Count), and a list of representative keywords. These topics reflect common patterns in the data.

B. Similarity Based Social Network

The social network was constructed to represent students as vertices and their relationships as edges based on their word pattern similarities. This process utilized the topic probabilities generated from the BERTopic model and relied on cosine similarity between the embedding vectors to quantify the relationships between student responses.

Given a set of student responses $D = \{d_1, d_2, \ldots, d_n\}$, each response d_i is encoded into a topic probability vector

$$\mathbf{p}_i = \{P(T_1 \mid d_i), P(T_2 \mid d_i), \ldots, P(T_k \mid d_i)\}$$

where $P(T_j \mid d_i)$ represents the probability of d_i belonging to topic T_j, and k is the number of topics.

To compute the relationships between responses, the cosine similarity $\text{Sim}(d_i, d_j)$ between the topic probability vectors $\mathbf{p}_i$ and $\mathbf{p}_j$ of two responses was calculated as:

$$\text{Sim}(d_i, d_j) = \frac{\mathbf{p}_i \cdot \mathbf{p}_j}{\|\mathbf{p}_i\| \|\mathbf{p}_j\|}$$

where $\|\mathbf{p}_i\|$ denotes the Euclidean norm of $\mathbf{p}_i$. The similarity score measures the alignment between two responses in terms of their topic distributions.

The graph $G = (V, E)$ is then constructed as follows:

- Each node $v_i \in V$ corresponds to a student where t_i is the most probable topic for d_i, determined as: $t_i = \text{argmax}_j \ P(T_j \mid d_i)$.
- An edge $(v_i, v_j) \in E$ is added between two nodes if their cosine similarity exceeds a predefined threshold τ:

$$(v_i, v_j) \in E \iff \text{Sim}(d_i, d_j) > \tau, \quad \tau = 0.90$$

- In BERTopic, the topic indices start from -1, indicating that topic -1 is an outlier that has very few similar documents. As some of the documents have a similarity less than threshold τ, they are assigned to the sub-community having the closest similarity to the topic -1.

The edge weight w_{ij} is set to the cosine similarity value: $w_{ij} = \text{Sim}(d_i, d_j)$. An example social network formed by student response data is shown in the Figure 1.

C. Community Detection

Community detection is a crucial step in our analysis for identifying groups of students with shared characteristics. We used the Louvain method and a hierarchical extension to discover primary communities and sub-communities within the student social network. The Louvain method is an iterative, modularity-based algorithm for detecting communities in a graph $G = (V, E)$, where V represents the set of nodes (students) and E represents the edges (relationships) between them. Modularity Q is used as the optimization objective to evaluate the quality of a community partition, defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where, A_{ij} is the weight of the edge between nodes i and j, k_i is the sum of the edge weights connected to node i, m is the total weight of all edges in the graph, c_i is the community assignment of node i, and $\delta(c_i, c_j)$ is an indicator function, equal to 1 if $c_i = c_j$, and 0 otherwise.

The Louvain algorithm proceeds in two phases:

1) **Local Modularity Optimization:** Each node is initially assigned to its own community. Nodes are moved between communities to maximize the modularity Q.
2) **Community Aggregation:** Communities identified in the first phase are collapsed into super-nodes, creating a new, smaller graph. The process is repeated until no further modularity improvement is possible.

The output is a partition $P = \{C_1, C_2, \ldots, C_k\}$, where each C_i is a community.

To capture fine-grained communities, we applied a recent extension to the Louvain method recursively where, after detecting the primary communities P, we apply the Louvain method to each community's subgraph [23].

1) **Primary Community Detection:** The Louvain method is applied to the original graph G to obtain primary communities:

$$P = \text{Louvain}(G)$$

Each community $C_i \in P$ forms a subgraph G_i:

$$G_i = G[V_i], \quad V_i = \{v \in V \mid v \in C_i\}$$

2) **Sub-Community Detection:** The Louvain method is recursively applied to each subgraph G_i to uncover sub-communities: $P_i = \text{Louvain}(G_i)$. This dynamic process enables the identification of nested community structures within the graph.
3) **Filtering Sub-Communities:** We have used modularity gain to detect sub-communities in the main community. For a community partition to be meaningful, its modularity score should improve significantly when divided into sub-communities. We define "sufficient modularity gain" as the improvement in modularity $\Delta Q = Q_{\text{new}} - Q_{current}$ between the current level of the graph and its sub-communities, where $Q_{current}$ is the modularity of the current partition and Q_{new} is the modularity of the sub-partition after further dividing the community. If ΔQ is below a predefined threshold ϵ (e.g., $\epsilon = 0.5$), the division stops.

Figure 1 shows the results of Hierarchical Louvain applied to the student social network where the main communities are represented by different colors. We also extract the subcommuntiies; however, they are not able to be shown in this figure but are captured in Table II.

D. Course Ranking

The course ranking and selection procedure leverages Jensen-Shannon Divergence (JSD) to determine the similarity between a new student's topic distribution and those of students within the identified sub-community. The similarity scores are combined with grade point information to rank courses taken by students in these communities and recommend the top options.

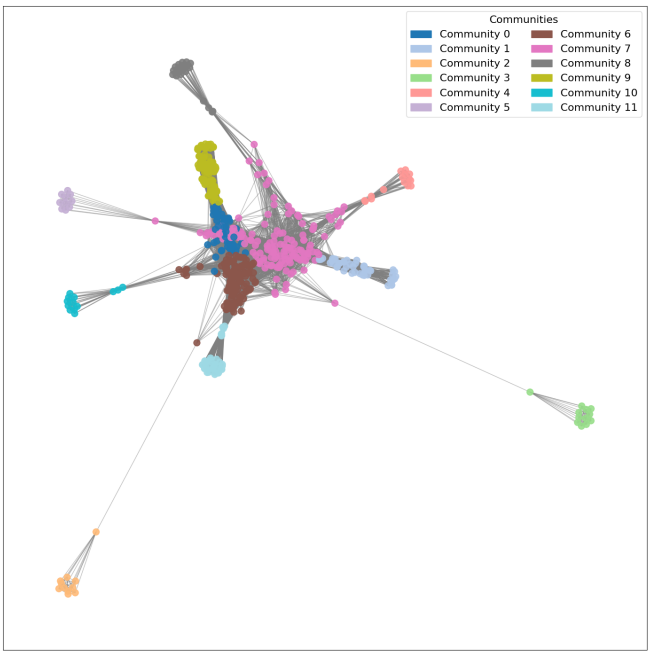


Fig. 1: Hierarchical Louvain applied to student network

Probability Normalization: The topic probability vector for the new student's response, $\mathbf{p}_{\text{new}}$, is normalized to ensure it sums to 1:

$$\mathbf{p}_{\text{new}} = \frac{\mathbf{p}_{\text{new}}}{\sum_{j=1}^{k} p_{\text{new},j}}$$

where $p_{\text{new},j}$ represents the probability of the new student's response belonging to topic j, and k is the total number of topics. Similarly, the topic probability vectors for each student in the sub-community, $\mathbf{p}_i$, are normalized:

$$\mathbf{p}_i = \frac{\mathbf{p}_i}{\sum_{j=1}^{k} p_{i,j}}$$

Jensen-Shannon Divergence Calculation: The Jensen-Shannon Divergence (JSD) is used to measure the similarity between the new student's normalized topic distribution $\mathbf{p}_{\text{new}}$ and each sub-community member's distribution $\mathbf{p}_i$. JSD is defined as:

$$\text{JSD}(\mathbf{p}_{\text{new}}, \mathbf{p}_i) = \frac{1}{2} \sum_{j=1}^{k} \left[p_{\text{new},j} \log \frac{p_{\text{new},j}}{m_j} + p_{i,j} \log \frac{p_{i,j}}{m_j} \right]$$

where:

$$m_j = \frac{p_{\text{new},j} + p_{i,j}}{2}$$

JSD ranges from 0 to 1, with smaller values indicating greater similarity between the two distributions.

To convert the divergence into a similarity score, let

$$\text{Sim}_i = \max(0, 1 - \text{JSD}(\mathbf{p}_{\text{new}}, \mathbf{p}_i))$$

This bounds the similarity score to the range $[0, 1]$.

Normalizing Similarity Scores: The similarity scores for all sub-community members are normalized to ensure their sum equals 1:

$$\text{Sim}_i^{\text{norm}} = \frac{\text{Sim}_i}{\sum_{i=1}^{n} \text{Sim}_i}, \quad \text{if} \quad \sum_{i=1}^{n} \text{Sim}_i > 0$$

where n is the number of sub-community members.

Course Weight Calculation: For calculating the course weight, JSD similarity and grade points are aggregated for all students in that particular subject. For a course c, the weighted similarity score is calculated as:

$$\text{WeightedScore}_c = \frac{\sum_{s \in S_c} (\text{Sim}_s \times G_s)}{|S_c|}$$

where:

- S_c is the set of students who have taken course c,
- Sim_s is the similarity score for student s,
- G_s is the grade points achieved by student s in course c,
- $|S_c|$ is the total number of students who have taken the course.

This approach computes the average weighted score for each course. This way we can ensure that courses taken by similar students with high grades receive higher scores.

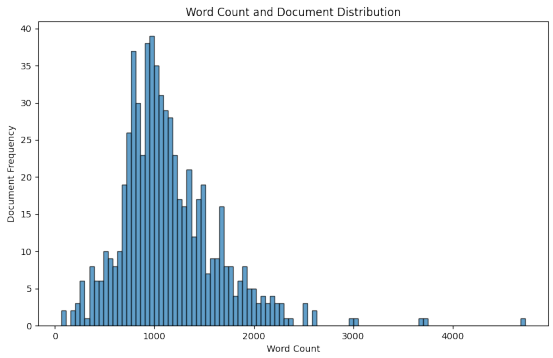
Course Ranking: The weighted scores for all courses are aggregated, and the top N courses (e.g., $N = 5$) are selected for recommendation. This procedure ranks courses based on their relevance to the new student, as determined by both textual similarity (captured through topic probabilities) and academic performance (captured through grades).

VI. Experimental Design

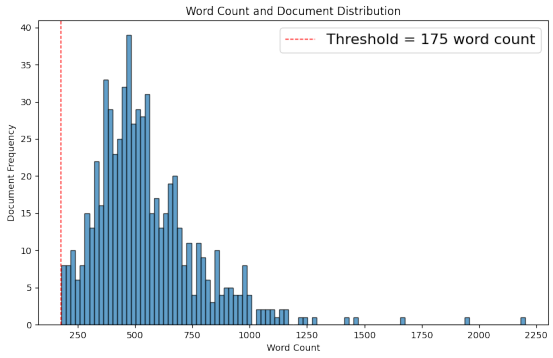
While the following does not provide a comprehensive evaluation of our approach, we discuss a sample application of the approach to real students at our university. To implement our approach, we constructed the topic model using the student responses. In the initial step, we utilized Python's NLTK library [24] to preprocess the text-based data by performing tasks such as removing digits, punctuation, and stop words, as well as lemmatizing and tokenizing words. Figure 2a shows the word count distribution across documents before preprocessing. We can see that the ESSAYS dataset has a wide range of word counts, including very short responses. Figure 2b, shows the word count after lemmatization and filtering. We excluded documents with fewer than 175 words, as such short responses were deemed insufficient for meaningful analysis.

The ESSAYS dataset contains responses for the Hilleman Scholars program from the years 2016 – 2024. Each year the number of questionnaires and the questions asked change. To address the problem, we have combined answers for all of the responses provided by each student into one contiguous document. This document was then used as the basis for generating the topics.

After preprocessing and lemmatization, we fine-tuned the BERTopic model using the written responses. We used the pre-trained paraphrase-mpnet-base-v2 model from the SentenceTransformer library [25] as the embedding model. This model is designed for semantic similarity tasks and works well with unstructured textual data. It creates dense embeddings that represent the semantic structure of the responses. For dimensionality reduction, we used UMAP with parameters that balance local and global structures. We set the neighborhood size to 15, which helps the embeddings capture both local relationships and global patterns. The embeddings

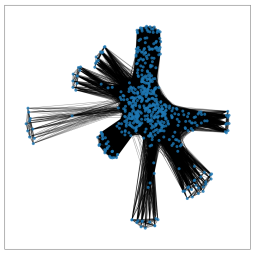


(a) Word distribution before lemmatization

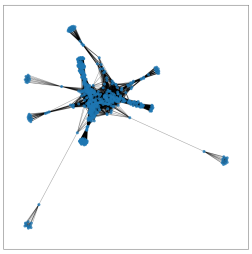


(b) Word distribution after lemmatization

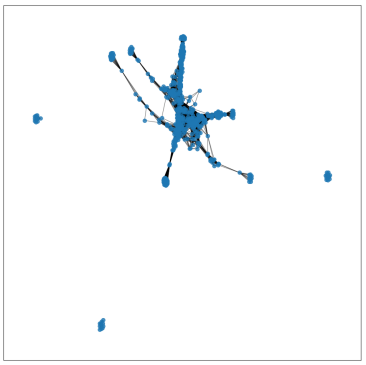
Fig. 2: Word frequency distribution by document



(a) 60% similarity threshold



(b) 90% similarity threshold



(c) 95% similarity threshold

Fig. 3: Social network for different similarity thresholds

were reduced to 5 dimensions to make computations faster and remove noise.

To create compact clusters, we set the minimum distance to 0.0. Cosine similarity was chosen as the distance metric because it handles high-dimensional embeddings effectively. For clustering, we used HDBSCAN with a minimum cluster size of 10. This ensures that each topic had at least 10 documents and prevented overly small or insignificant topics. Although we used cosine similarity for generating embeddings, Euclidean distance was used for clustering, as it works well with UMAP-reduced embeddings. We used the "excess of mass" method [26] to find dense regions in the data in order to identify coherent topics. We also configured BERTopic to improve topic extraction. A minimum topic size of 10 ensured that small and irrelevant topics were avoided. The model also provides topic probabilities for each document, which gives detailed insights into the word patterns in the responses.

Since this was a simulated experiment, we used data based on students already admitted to consider recommendations vs courses taken. For this experiment, the combined dataset was split into 90% training data and 10% test data. We recognize that the test set is small, but the dataset itself is limited. Allocating too many responses to the test set would create smaller and less coherent social groups. To construct the social network, we began by encoding the student responses into dense vector representations using a pre-trained embedding model. The graph was then initialized using the networkx library [27]. Each node consist of two attributes: their corresponding embedding and topic assignment. For weighting edge between nodes, we computed the pairwise cosine similarity of their topic probability vectors. An edge was added between two nodes if their similarity exceeded a tuned threshold of 90%. When a very low threshold was used, (e.g., 60% as in Figure 3a), the network became overly dense, resulting in decreased modularity. In this case, the Louvain algorithm struggles to identify meaningful communities because the excessive number of connections obscures the underlying structure. On the other hand, when a very high threshold, such as 95%, was applied (Figure 3c), the network became a forest with many disconnected components. This reduces edge density, and the Louvain algorithm has difficulty optimizing modularity due to the lack of sufficient connections. We found the 90% similarity threshold to be optimal.

VII. Results and Discussion

Table II shows the results of applying Hierarchical Louvain to our student network. It detected a total of 12 main communities within the network, and each main community was divided further into sub-communities based on modularity optimization. From the results, we can see that the size and structure of the communities vary significantly. Main Community 0, for example, is divided into three sub-communities with 26, 19, and 9 nodes, respectively. On the other hand, the largest community, Main Community 7 has four sub-communities containing 58, 44, 30, and 22 nodes.

TABLE II: Communities and sub-communities discovered

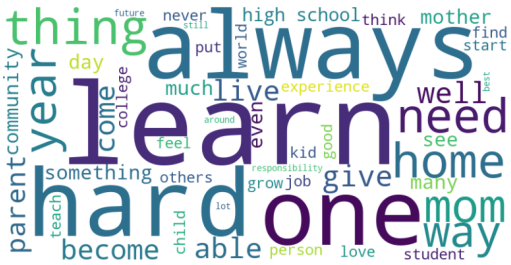
Main Community	Sub-Community	Number of Students
0	Sub-Community 0	26
	Sub-Community 1	19
	Sub-Community 2	9
1	Sub-Community 0	22
	Sub-Community 1	24
2	Sub-Community 0	11
3	Sub-Community 0	12
4	Sub-Community 0	15
5	Sub-Community 0	10
6	Sub-Community 0	57
	Sub-Community 1	60
7	Sub-Community 0	30
	Sub-Community 1	58
	Sub-Community 2	22
	Sub-Community 3	44
8	Sub-Community 0	18
9	Sub-Community 0	53
	Sub-Community 1	53
10	Sub-Community 0	13
11	Sub-Community 0	43

From this, we deduce that NH-Louvain can adapt to the density and size of different network regions and can capture fine-grained relationships. The smaller communities, such as occur in Main Communities 2, 3, 4, and 5, each contain a single sub-community (i.e., are not decomposed further) with fewer than 20 nodes, indicating more localized and isolated clusters.

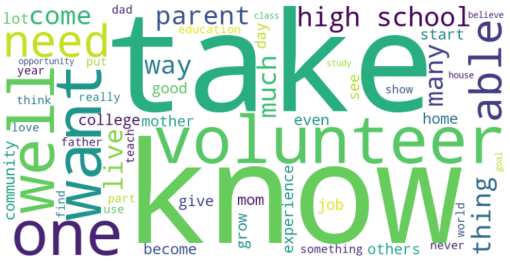
As shown in Figure 1, densely connected regions represent large sub-communities and sparse connections, indicate smaller or more isolated groups. This highlights the modular nature of the network, where students cluster based on shared interests, experiences, or similarities in their responses.

Figure 4 presents sample WordClouds for three of the largest discovered communities. These WordClouds highlight the most important differentiating terms based on differences in the topic distributions used by students within each community. The WordCloud in Figure 4a for Main Community 0 shows prominent terms such as “learn,” “always,” “need,” “hard,” and “home.” This suggests a strong focus on personal development, resilience, and education. The WordCloud in Figure 4b for Main Community 6 includes key terms such as “know,” “volunteer,” “take,” and “want.” This community has a similar self-development and education theme but has a distinct emphasis on volunteering and community involvement. The WordCloud in Figure 4c for Main Community 7 highlights terms like “able,” “need,” “one,” and “become.” This community appears to focus on personal capability, transformation, and ethical values, as suggested by words like “ethic” and “good.”

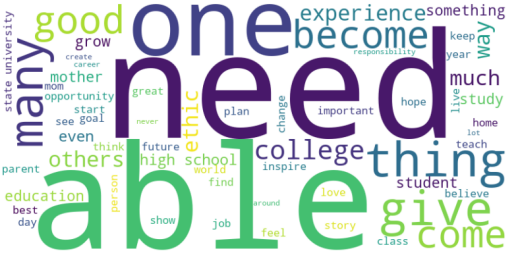
After discovering communities and sub-communites, we tested our method with the 10% test data. We randomly selected a Student ID from the test set, calculated its probability distribution and matched it with the nearest node in the social network. After that we queried our database with the Student IDs of the best matched community, determined to be the community to which that nearest node belonged, and



(a) Main community 0



(b) Main community 6



(c) Main community 7

Fig. 4: Word clouds of the three largest communities

ranked the core courses outlined in Section V-D to select the N most appropriate courses for the student. Table III shows the ranked courses recommended for the student by our recommender system.

When examining this table, we see that a wide variety of courses are being recommended to the student, ranked based on other students’ prior performance, combined with the common interests of these students, as reflected in the topic model-generated social network. In some cases, multiple choices in the same category are suggested, such as LIT 169IH and JPNS 325IH, both of which are humanities core courses. Even, so the literature courses is ranked higher than the Japanese studies course, perhaps due to the student’s specific areas of interest.

Note that some students may have already completed core courses, in which case, such recommendations should be filtered out. We did not include that filter in this demonstration. In addition, some courses have conditions that must be satisfied before the students is eligible to take it. For example, HONR 201IH is a “humanities” core course offered as part

TABLE III: Recommended core courses for a sample student

Course ID	Course Name	Avg. Weight
HONR 202IH	Text & Critics	4.00
ERTH 102CS	Topics in Earth Science	4.00
LIT 169IH	Literature as Popular Culture	3.85
SOCI 318R	Sociological Research Methods	3.70
GRMN 102D	German 2	3.57
SOCI 150D	Social Difference	3.30
M 171Q	Calculus 1	3.19
ARCH 121IA	Intro Design	3.06
JPNS 325IH	Outcast Literature	3.00

of the university's Honors College. The student would need to be a member of the Honors College to be eligible to take this course. Furthermore, GRMN 102D is the second course in the introductory German sequence, so the student would have needed to complete the equivalent of GRMN 101 before being able to enroll in this course. None of these constraints were considered in this initial demonstration of our method.

VIII. Conclusion and Future Work

In this paper we presented a new approach to recommend "core" courses for incoming students, especially those at a disadvantage, based on topics derived from their responses to a variety of general interest questions and their similarity with to topics associated with prior students. The idea was to match interests to courses that also have a history of successful student completion as a way of enhancing student success and promoting a sense of community, purpose, and belonging among the incoming students. We acknowledge that our methodology is still in the early stages, for example in how we are ranking the courses. As our data already include other student information, such as writing and math placement scores (i.e., standardized test scores), we anticipate that incorporating this information may result in more robust course recommendations.

The primary area of future work is a full-scale evaluation of the method on the effects of student retention. We also plan to develop explainability methods to provide better insight into the recommendations made. Finally, there is the potential for biased recommendations due to possible latent biases in the data, especially with respect to how topics are filtered.

While our work is exploratory, the practical and prospective application of our method allows universities improved means to support students through common/matched pair interests, improved sense of purpose and belonging, and justifiable course registrations. This is consistent with both emerging needs and long-standing practices in student development, where student groups might be organized as informed by the same or similar characteristics of their more tenured student peers.

References

[1] G. Kuh, J. Kinzie, Buckley, B. J.A, B.K, and J. Hayek, "Piecing together the student success puzzle: Research, propositions, and recommendations," *ASHE Higher Education Report*, vol. 32, no. 5, 2007.

[2] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[3] E. Atagün, B. Hartoka, and A. Albayrak, "Topic modeling using LDA and BERT techniques: Teknofest example," in *6th International Conference on Computer Science and Engineering (UBMK)*, 2021, pp. 660–664.

[4] A. K. Bhowmick, K. Meneni, M. Danisch, J.-L. Guillaume, and B. Mitra, "LouvainNE: Hierarchical Louvain method for high quality and scalable network embedding," in *Proceedings of the 13th ACM International Conference on Web Search and Data Mining*, 2020, pp. 43–51.

[5] N. Bendakir and E. Aïmeur, "Using association rules for course recommendation," in *AAAI Workshop on Educational Data Mining*, 2006, pp. 31–40.

[6] A. Elbadrawy and G. Karypis, "Domain-aware grade prediction and top-n course recommendation," in *RecSys '16: Tenth ACM Conference on Recommender Systems*, 2016, pp. 183–190.

[7] P. Chang, C. Lin, and M. Chen, "A hybrid course recommendation system by integrating collaborative filtering and artificial immune systems," *Algorithms*, vol. 9, no. 3, p. 47, 2016.

[8] G. Koutrika, B. Bercovitz, and H. Garcia-Molina, "Flexrecs: Expressing and combining flexible recommendations," in *SIGMOD/PODS '09: International Conference on Management of Data*, 2009, pp. 745–758.

[9] H. Zhang, T. Huang, Z. Lv, S. Liu, and Z. Zhou, "MCRS: A course recommendation system for moocs," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 7051–7069, Mar 2018.

[10] P. B. Thorat, R. M. Goudar, and S. Barve, "Survey on collaborative filtering, content-based filtering and hybrid recommendation system," *International Journal of Computer Applications*, vol. 110, no. 4, pp. 31–36, 2015.

[11] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1, pp. 177–196, Jan 2001.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. null, p. 993–1022, 2003.

[13] X. Jiang, L. Bai, X. Yan, and Y. Wang, "LDA-based online intelligent courses recommendation system," *Evolutionary Intelligence*, vol. 16, no. 5, pp. 1619–1625, Oct 2023.

[14] K. A. R. Riyadi, "Hybrid movie recommendation system using neural collaborative filtering and BERTopic," Ph.D. dissertation, Universitas Gadjah Mada, 2023.

[15] Y. Ng and J. Linn, "CrsRecs: A personalized course recommendation system," in *Proceedings of the International Conference on Advanced Computing*, 2022, pp. 89–95.

[16] Y. Yang and G. Yuan, "Oral English course online recommendation system based on support vector machine," in *Intelligent Computing Technology and Automation*. IOS Press, 2024, pp. 685–692.

[17] M. A. Noor, J. Sheppard, and J. Clark, "Finding potential research collaborations from social networks derived from topic models," in *2023 10th International Conference on Behavioural and Social Computing (BESC)*. IEEE, 2023, pp. 1–7.

[18] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, October 2008.

[19] M. A. Noor, J. A. Clark, and J. W. Sheppard, "ScholarNodes: Applying content-based filtering to recommend interdisciplinary communities within scholarly social networks," in *Proceedings of the 47th ACM International SIGIR Conference on Research and Development in Information Retrieval*. Washington, DC, USA: ACM, 2024, pp. 1–5.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[21] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[22] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[23] M. A. Noor, J. W. Sheppard, and J. A. Clark, "Identifying hierarchical community structures in content-based scholarly social networks," in *Proceedings of the IEEE International Conference on Machine Learning*, 2024.

[24] E. Loper and S. Bird, "NLTK: the natural language toolkit," *arXiv preprint arXiv:cs.CL/0205028*, 2002.

[25] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-Networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2019.

[26] L. McInnes, J. Healy, and S. Astels, "Parameter selection for hdbscan," Online, 2025, available at https://hdbscan.readthedocs.io/en/latest/parameter_selection.html.

[27] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conference*, August 2008, pp. 11–15.