Semantic Topic Chains for Modeling Temporality of Themes in Online Student Discussion Forums

Harshita Chopra
Adobe Research, India
harshitac@adobe.com

Yiwen Lin
University of California, Irvine
yiwenl21@uci.edu

Mohammad Amin Samadi
University of California, Irvine
masamadi@uci.edu

Jacqueline G. Cavazos
University of California, Irvine
jacqueline.cavazos@uci.edu

Renzhe Yu
Columbia University
renzheyu@tc.columbia.edu

Spencer Jaquay
University of California, Irvine
sjaquay@uci.edu

Nia Nixon
University of California, Irvine
dowelln@uci.edu

ABSTRACT

Exploring students' discourse in academic settings over time can provide valuable insight into the evolution of learner engagement and participation in online learning. In this study, we propose an analytical framework to capture topics and the temporal progression of learner discourse. We employed a Contextualized Topic Modeling technique on messages posted by undergraduates in online discussion forums from Fall 2019 to Spring 2020. We further evaluated if topics were originating from specific courses or more generally distributed across multiple courses. Our results suggested a significant increase in the number of general topics after the onset of the pandemic, suggesting emergent topics being discussed in a range of courses. In addition, using Word Mover's Distance, we examined the semantic similarity of topics in adjacent months and constructed topic chains. Our findings indicated that previously course-centric topics such as public health developed into more general discussions that emphasize inequities and healthcare during the pandemic. Furthermore, emergent topics around students' lived experiences underscored the role of discussion forums in capturing educational experiences temporally. Finally, we discuss the implications of current findings for post-pandemic higher education and the effectiveness of our framework in exploring unstructured large-scale educational data.

Keywords

Topic Modeling, Topic Detection and Tracking, Natural Language Processing, Discourse Analysis

1. INTRODUCTION

H. Chopra, Y. Lin, M. A. Samadi, J. G. Cavazos, R. Yu, S. Jaquay, and N. Nixon. Semantic topic chains for modeling temporality of themes in online student discussion forums. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 67–78, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115691

Social learning theories suggest that learning occurs through interaction with peers and instructors[52]. In fully online learning contexts, interpersonal interaction plays an even more important role and discussion forum is one of the most commonly used tool to enable interactions and facilitate classroom community. As such, examining text and language in a discussion forum allows us to gain insights on learning since language plays an instrumental role in promoting thinking and knowledge construction as well as social activities occurring in computer-mediated environments[51]. While the use of discussion forums has been prominent on informal learning platforms such as Massive Open Online Courses (MOOCs) [4], the adoption of discussion forums for social learning and communication across accredited institutions has been less systematic until the contingency shift to remote learning due to COVID-19.

In late 2019, news began to spread about a respiratory illness appearing throughout parts of the globe. By March 2020, the World Health Organization officially declared COVID-19 a global pandemic [57]. In response to the global health crisis, universities and schools quickly suspended in-person classes and shifted to fully online instruction[50]. Although online learning and teaching are already prevalent, the urgent shift to online education created a nontrivial disruption to students' educational experience[50]. During this unprecedented time, undergraduate students were confronted with new health concerns, challenges adapting to online learning, and seeking ways to build the relationships and connect with peers that are essential for students' college success. The outbreak has also posed challenges for instructors to ensure instructional delivery and quality through a variety of tools such as video conferencing, forum discussions and assignment submission through Learning Management Systems (LMS).

As learners were prompted to fully remote learning, the amount of educational data generated day by day became unprecedented. Since the pandemic was a major disruption to regular instruction, the shift to fully remote instruction derived a need to harvest insight into how learning activities were impacted [33]. Recent advances in data mining tech-

niques provide efficient and promising tools to effectively process large-scale educational data to model learner behavior in real-time [1]. Moreover, leveraging objective indicators rather than self-report data allows a non-intrusive way to explore the consequences of the transition to online learning and the impact of COVID-19 on the educational process [11]. Digital trace data generated from learner's activities in LMS are not only useful in tracking engagement patterns and predicting learning outcomes across informal and formal education context [30], but features such as textual data also allow for non-intrusive means to analyze social and cognitive dynamics compared to interviews or questionnaire for data sampling [46]. The applications of Natural Language Processing (NLP) techniques in education have been proven to be powerful and effective tools for modeling learner behavior and extracting meaningful information to enhance our understanding of social interactions in computer-mediated learning environments [20, 32, 17, 19, 22]. A recent systematic review also identifies textual analysis of asynchronous discussion forums arises as an emergent theme in learning analytics given the increasing use of online LMS platforms[2]. This suggests that increased attention in the field of learning analytics has been given to exploratory approaches to harvesting insights from educational discourse.

Given the COVID-19 circumstances and the constantly changing nature of the pandemic and policy responses, automated methods focused on detecting the temporal aspect of learning activities could be valuable in providing insights to policymakers and instructors on the overall changing landscape in online learning. Indeed, text mining techniques offer a viable way to extract meaningful information and unravel latent patterns from large-scale educational data. Taken together, we built on our previous study [15] to propose an analytical framework that characterizes the evolution of topics in unstructured student discourse present in online discussion forums. We applied this framework to derive a series of semantically meaningful topic chains that we believe reflect the changes in online learning activities due to the emergency shift to remote learning and the COVID-19 pandemic influence.

This paper employs novel NLP approaches to explore the emergent topics in LMS and reveal the temporal development of teaching and learning activities in online discussion forums. Specifically, we investigate the topics before and after universities transitioned to a fully remote format to reflect the organic responses from teachers and students to this significant disruption in teaching and learning. As a methodological contribution, our framework combines state-of-the-art topic modeling techniques (i.e., Contextualized Topic Modeling and Word Mover Distance) to construct semantically relevant topics on a month-to-month granularity. This framework offers an opportunity to examine topical evolution in educational data. Despite the popularity of topic modeling in social media research, it remains a niche cluster in educational research [2] and focuses on static content classification[25]. In contrast, the temporal dynamics of topics highlighted in our framework would be a unique contribution in this area. Additionally, we extracted course centrality as a course-level feature to further contextualize topic chains. This allows us to distinguish topics that are specific to a discipline from more generally occurring topics in the entire forum environment. Finally, our study presents an exploratory rather than prescriptive view based on emergent student discourse revealing the impact of larger societal events on educational activities and a transparent analytic process to harvest information that aids decision-making.

The rest of the paper is organized as follows. First, we provide a brief overview of text mining practices in education in recent years. Second, we describe specific topic modeling techniques for processing large-scale text corpus. We then introduce our research questions and the methodological approach of the current study. Finally, we present our findings with a discussion on the implications of our study and the potential adaptation of this analytical framework for the broader education data mining community.

2. BACKGROUND

2.1 Text Mining in Education

Language is a channel for expressing people's opinions and experiences. The advances in computational linguistics offer ways to systematically explore these experiences and capture trends in discourse through individuals' language use. In educational contexts, the exponential growth of learner data generated in the digital environment has prompted the need to apply data-driven approaches to handle and make sense of learner data at scale [26]. Student essays, online forums, and online assignments are major venues and resources for large-scale text mining analysis to derive informative insights for evaluating performance or providing analytics insights for instructors and students to support learning[25]. NLP and machine learning algorithms are promising tools to automate this process. Previous studies have assessed discourse using text summarization [27], examined sociocognitive processes in collaborative conversations [21, 46], and modeled learner trajectories using neural network-based predictive methods [13].

Amongst an array of NLP techniques, topic modeling is an unsupervised method to extract emergent thematic structure in large textual data by connecting documents that share similar patterns[3, 34]. Topic modeling is deemed a powerful tool for education data mining and learning analytics research[48]. In empirical studies, topic modeling has been applied to examine course reviews to extract learner interests in order to provide personalized course recommendations to improve the quality of online courses[38, 42], to identify themes in asynchronous online discussions for providing adaptive support to individual students[23], or characterize chat dialogues within learning groups to support collaborative learning[12], or evaluating students' reflective writing to examine their learning experiences and engagement associated with course content[14].

Previously, research has pointed out discussion forums as an under-explored territory, where rich textual dialogue could offer the potential to support student learning[23]. While discussion forums may appear to be plagued by information overload and chaos, applying appropriate analytic techniques can help bring structure and identify important threads for students and teachers[56]. A growing body of learning analytics research focuses on discussion forum data to measure learner participation, and examine the associ-

ations between discourse behavior and learning[54]. Some studies leverage linguistic features such as sentiment valence to predict learning gains and retention for individual learners [55][37]. Other studies took a more exploratory perspective to investigate the temporal changes in MOOC learners' language and discourse characteristics at a high level [18]. As we have seen online discussion forums delivering promising results for social learning and student-teacher interactions in MOOCs[10], a question arises when a large number of courses at accredited universities started to rely on discussion forums during the pandemic: How can we leverage the large-scale, extensive data that emerged in the months before and after the outbreak of COVID-19 to understand discussion forum activities better? In fully remote instructions under the pandemic influence, the discussion forum is considered one of the most commonly used tools for supporting social interactions within online courses to meet the needs of a diverse student population[41]. Therefore, getting a better picture of the emergent topic and trends in this instructional environment would be beneficial to understanding how teaching and learning were affected during the critical months of the pandemic impact.

In investigating the impact of COVID-19, topic modeling has shown effectiveness in reflecting trends and public opinions on social media and online responses towards policy decisions[9, 47]. [39] studied the online mental health support community on Reddit (e.g., r/HealthAnxiety, r/schizophrenia) during the COVID-19 pandemic, looking into how different patterns of behavior can be captured through language and topic modeling. The increased use of discussion forum during the pandemic presents an opportunity to investigate learning activity in formal educational spaces. In particular, the timeline of COVID-19 spread and the academic calendar creates an unique alignment for potential changes in discussion forum discourse due to the influence of the pandemic and the transition to fully online learning. However, to our knowledge, no study has considered constructing temporal chains of topics in learner discourse that tracks these changes. Therefore, we seek to examine the rich textual data that exists in the Learning Management System (LMS) before and after universities shifted to remote learning.

2.2 Temporal Topic Modeling

With the ever-growing rate of data generation, topic modeling is a widely used approach to find patterns and trends within large-scale non-annotated data. Among several topic modeling techniques in NLP [34], one of the most widely used models is Latent Dirichlet Allocation (LDA) [8]. LDA is a generative statistical model that is used to extract latent topics from a text corpus. LDA models document as a probability distribution over topics where each topic is a probability distribution over words in the vocabulary. Recent advances in deep learning have introduced the combination of neural networks and transformer-based techniques [58, 28] to yield topics that are more coherent and interpretable than traditional models that use only bag-of-words (BoW) features. For instance, Combined Topic Model (CombinedTM) [5] is a recently proposed neural topic model that combines the bag of words (BoW) approach and the neural topic model ProdLDA [49] with the contextualized document representations from Sentence-BERT [45] for more coherent topics. CombinedTM is a Contextualized Topic Model (CTM) which uses Bag of Words (BoW) document representation concatenated with the contextualized document representations from Sentence-BERT [45] converted to the same dimensionality as of the BoW vocabulary by using a hidden neural network layer. This latent representation of the document is passed through a decoder network that reconstructs the BoW. The framework is originally based on a variational inference process [43]. Compared to existing topic models including the traditional LDA, CombinedTM showed more coherent topics and added contextual information [5]. In this study, we leverage CombinedTM to explore emerging topics and their evolution in learners' discussion forums before and during the onset of the COVID-19 pandemic.

Exploring how topics evolve can help us discover how certain events, such as the global health crisis, remote learning, and social injustice, impact learners' lived and academic experiences and shape their learning activities during the pandemic. Previous studies have proposed Dynamic Topic Model (DTM) [7] and related models [53] to detect and track topics over time-sequenced texts. However, these frameworks only model the variation in the probability distribution of words within a constant set of topics across time. To track how independent topics emerge, decline, and shift focus in a sequentially sliced corpus, researchers have used topic models and similarity metrics to connect topics in adjacent time steps [35, 31]. Similarity scores, such as Jaccard Coefficient, used in these studies do not account for the semantic similarity between words appearing in related contexts (e.g., college and university). To address this limitation, we propose a framework to detect and track semantically similar topics using WMD [36]. WMD measures the dissimilarity between two text documents, leveraging the power of word embeddings trained on a text corpus using the Word2Vec model [40]. We use this approach to find topics that represent a similar broad theme but depict a change in context in subsequent time steps. Notably, the WMD between two documents can be computed in a meaningful way even if the two documents do not have any words in common. WMD does not consider the order of words, making it suitable for tracking the similarity between two sets of words representing topics.

3. CURRENT STUDY

The objective of this study is to explore the emergent topics and their evolution in undergraduates' online discussion forums in the months prior to and after contingency shifts to fully remote learning due to the COVID-19 pandemic. Specifically, we aim to address the following questions:

> RQ1.a. What are the topics discussed by students before and after the pandemic?
> RQ1.b. How do these topics connect and evolve across months?
>
> RQ2: Do these topics represent discussions that are specific to certain courses or reflect general discourse across multiple courses?

To address the first research question, we used NLP techniques, specifically topic modeling, to capture the temporal

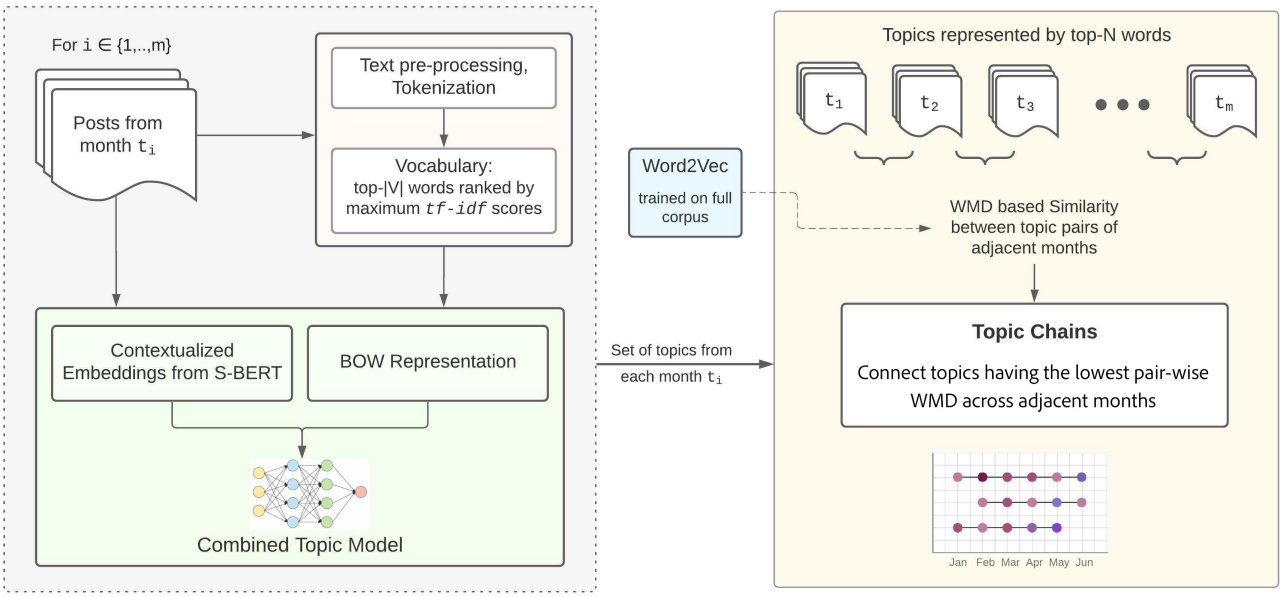

Figure 1: Graphical representation of the framework for topic modeling and topic evolution.

characteristics of learner discourse during this critical time in order to enhance our understanding of the influence of the pandemic on learning activities and learner discourse in a formal education setting. To achieve this goal, we introduce an analytical framework that comprises a CombinedTM [5] and a contextualized topic modeling (CTM) technique for extracting coherent themes that emerged in the discussion forum of the LMS across nine months. We then use Word Mover's Distance (WMD) to construct the linkage and temporal nuances of topics across months from October 2019 to June 2020 in order to build semantically meaningful topic chains that illustrate topical evolution over time. To address the second research question, we added a course-centricity measure to further contextualize whether a topic is specific to a course or discipline domain or more generic in the discussion forum.

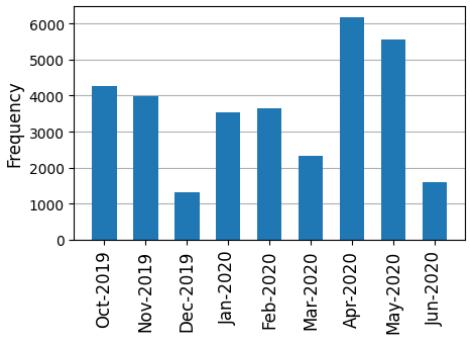

Figure 2: Frequency of posts in each month from October 2019 to June 2020.

Our current study contributes to the field of learning analytics by offering an analytic framework that is more interpretable for modeling topics in large-scale discourse data. The overview of our proposed framework is displayed in Fig. 1. We aim to extend the current literature by moving from mining static topics from large-scale student discourse to a more process-oriented perspective that shows how topics emerge, recur, and evolve over time. This analysis would allow us to view learner discourse as dynamic rather than a static picture and uncover relational linkages in the discourse.

4. DATA AND PARTICIPANTS

4.1 Participants

Our data were obtained from a large public university in the United States, which operates on a quarter system. We retrieved all discussion forum posts in the LMS across all courses offered in Fall 2019, Winter 2020, and Spring 2020 quarters (a total time span of nine months). We filtered the dataset to retain posts from a common set of students across each month (i.e., students who consistently contributed to the forum discussion). To eliminate individual differences, we focused on trends and themes of discourse originating from the same individuals. In addition, we considered posts that contained less than two words or had a length of fewer than five characters uninformative and removed them from the dataset. A total of 32,409 posts created by 449 students across 636 courses were retrieved, along with the time stamp of each post and the associated course and academic term. Given the rapid evolution of the pandemic, we use a month as the unit of analysis instead of the academic quarter to obtain more granular information about the temporal changes in student discourse. Fig. 2 shows the frequency of posts

by month. We observe that term structure impacted the frequency of messages posted in discussion forums. Conclusion of courses before academic breaks, for example, winter break (Dec 2019) and summer break (Jun 2020), led to a lesser engagement in online discussions.

Discussion posts include conversations from 310 female students and 139 male students who identified with the following ethnic backgrounds: Asian/Asian American (49.7%), Hispanic (29.2%), White Non-Hispanic (13.6%), Black (4.0%), American Indian/Alaskan Native (0.2%), and others (3.3%). Overall, 52.5% of students identified as first-generation college students.

4.2 Pre-processing

To prepare the text for analysis, we pre-processed the data using Natural Language Toolkit (NLTK) [6] and Gensim [44], two open-source libraries in Python3. Website links and email addresses were replaced with "url" and "email" tokens, respectively. Contractions were fixed, and irrelevant and redundant terms such as punctuation, digits, and stopwords were removed. To retain relevant words for building topic models, we used the Part-of-Speech Tagger to identify nouns, verbs, adjectives, and adverbs, which were further lemmatized using the WordNetLemmatizer from the NLTK package. Lemmatization retains the base word known as lemma (e.g., study) from inflected forms of a word (e.g., studying, studies, studied).

5. TOPIC DETECTION AND TRACKING

5.1 Topic Detection

Regarding RQ1.a, we first detected the emerging topics in the student's discussion forum data. To achieve this, we trained CombinedTM on the messages posted by students over each month separately. The quality and interpretability of the resulting topics depend on the curated vocabulary [24]. We constructed the BoW vocabulary by retrieving the top 10,000 words with maximum TF-IDF (Term Frequency - Inverse Document Frequency) weights. Sentence-BERT [45], a modification of the BERT [16] model, was used to obtain meaningful encodings of the messages. To optimize the number of topics (K) as a hyperparameter for each month, we ran the model on the corpus of each month with K ranging from 5 to 15 topics heuristically. While having less than five topics would result in overly generic and less coherent topics, on the other hand, a large number of topics causes highly coherent topics with lower topic diversity. Therefore, to optimize the number of topics, it is necessary to evaluate them on the three metrics used in CombinedTM that evaluate topic coherence and topic diversity:

1. Normalized pointwise mutual information
2. Word embeddings-based similarity
3. Inversed Rank-Biased Overlap

Normalized pointwise mutual information (NPMI) was used to measure how related the top-10 words of a topic are to each other, considering the words' empirical frequency in the original corpus. Word embeddings-based similarity refers to the average pairwise cosine similarity of the word embeddings of the top-10 words in a topic, using Word2Vec [40] word embeddings. The overall average of those values for all the topics was computed. Inversed Rank-Biased Overlap (IRBO) was used to score the diversity of the resulting topics, and NPMI and Word embeddings-based similarity were used to measure the average topic coherence. Subsequently, we used the soft voting ensemble approach to select the optimal K corresponding to the model which returned the highest sum of the normalized values of these metrics.

5.2 Topic Evolution

Previous research on topic tracking uses similarity metrics such as Kullback-Leibler (KL) divergence, Jaccard Coefficient, Kendall's Coefficient, and Cosine similarity between the probability distribution vectors of two topics. However, these metrics do not account for the semantic similarity between words of the same theme (e.g., student and university). Hence, in order to measure the semantic and contextual similarity between topics, we use the Word Mover's Distance (WMD). If there is a connection between two topics in consequent months, we consider that as an evolution of a topic and call it a "Topic Chain". More specifically, to address RQ1.b, we used the WMD to measure the semantic similarity between two topics. WMD finds an optimal solution to a transportation problem, which determines the minimum cost to move all words from one document to another. We trained a Word2Vec model using the skip-gram algorithm over the discussion posts data with a sliding window size of five, to obtain a lower-dimensional representation ($d = 100$) of words present in the entire corpus. Next, we represented each topic as a list of top-30 most frequent words in that topic and computed the WMD between two topics ϕ_i^t and ϕ_j^{t+1} for every i, j where $i \in \{0, ..., K_t\}$, $j \in \{0, ..., K_{t+1}\}$ and $t, t+1$ represent two consecutive months. For every topic ϕ_i^t in month t, we selected the topic from month $t+1$ which had the least WMD from it, thereby denoting the highest similarity. To keep a record of the connected topics, we created a mapping from ϕ_i^t to ϕ_j^{t+1} and saved the corresponding WMD. To avoid multiple topics in month t being mapped to the same topic in month $t+1$, we retained only the topic pairs having the least WMD among them. We created a directed graph connecting nodes (or topics) in consecutive months. Finally, we found all the simple paths from each root to leaf in this graph using the NetworkX package [29]. These directed paths are referred to as "topic chains". Effectively, we identified the most semantically similar topic pairs in subsequent months and connected them to construct topic chains.

5.3 Course-centricity of Topics

As noted above, discussion posts were taken from over 600 different courses. Inevitably, the content of each post can be heavily influenced by the course. Some discussions may be prompted by the course instructor or students, which as a result could impact the discussion forum conversations and the organic flow of the discussions. Consequently, these course-specific combined with several related courses could result in detected topics centered around course-related material. By contrast, other topics might represent a theme of discussions that originate from and are present in a broader range of courses, which makes the topic less dependent on course material resulting in more "general" topics. To address RQ2, we propose a measure of course-centricity to distinguish topics that are more specific to certain courses from topics that represent a broader theme that emerged repeatedly across mul-

Figure 3: Word-clouds of selected topics. The size of words is directly proportional to the topic-term probability, where a larger size represents higher relevancy to the topic.

tiple courses. We associated the course-centricity of a topic with the standard deviation of the distribution of message counts across courses. For each discussion post, we selected the topic with the highest probability of being assigned as the main topic of the post. Next, we selected the top ten courses with the most number of posts in each topic. The frequencies of the posts corresponding to the top-$N(= 10)$ most common courses were used to calculate the standard deviation (σ) of the distribution of posts for each topic. A lower value of σ denoted a relatively uniform distribution of courses in a topic, suggesting a topic is more generally distributed across courses. A higher value of σ denoted a skewed distribution where very few courses dominate the discussion, showing that the topic is more "course-centric". In addition, we measured the number of messages in each of the top ten courses for the topic. Topics that were mostly course-centric would have a disproportionately higher message count from one or two courses than from other courses, while more general topics would have a more even distribution of message count across the top ten courses.

6. RESULTS AND DISCUSSION

6.1 Topic Detection

The topic modeling resulted in 8-13 topics per month, including literature, philosophy, global health, and COVID-19-related discourse. We also observed that casual discourse and discussions on academic work remained the most common topics across all months. This implies that discussion forum participation includes not only learners' active knowledge construction that correlates with grades and learning outcomes, but also social connectedness that might affect learning experience online. Topics such as immigration and ethnic diversity, student lived experiences, social effects of technology and socializing in school revealed a deeper insight into students' personal views. Moreover, social justice movements such as the Black Lives Matter movement during the spring of 2020 emerged as a novel topic in discussion and conversation among students. This signals the influence of contemporary events on learning, and that students are actively trying to make sense of what is happening in the world and integrating that reflection into learning. Fig. 3 demonstrates the word clouds of selected topics discussed below. The topics are inferred by the authors from the distribution of words and posts representing them. A list of top-ranked words representing all the identified topics and their corresponding labels interpreted by the authors are publicly available[1].

[1]github.com/The-Language-and-Learning-Analytics-Lab

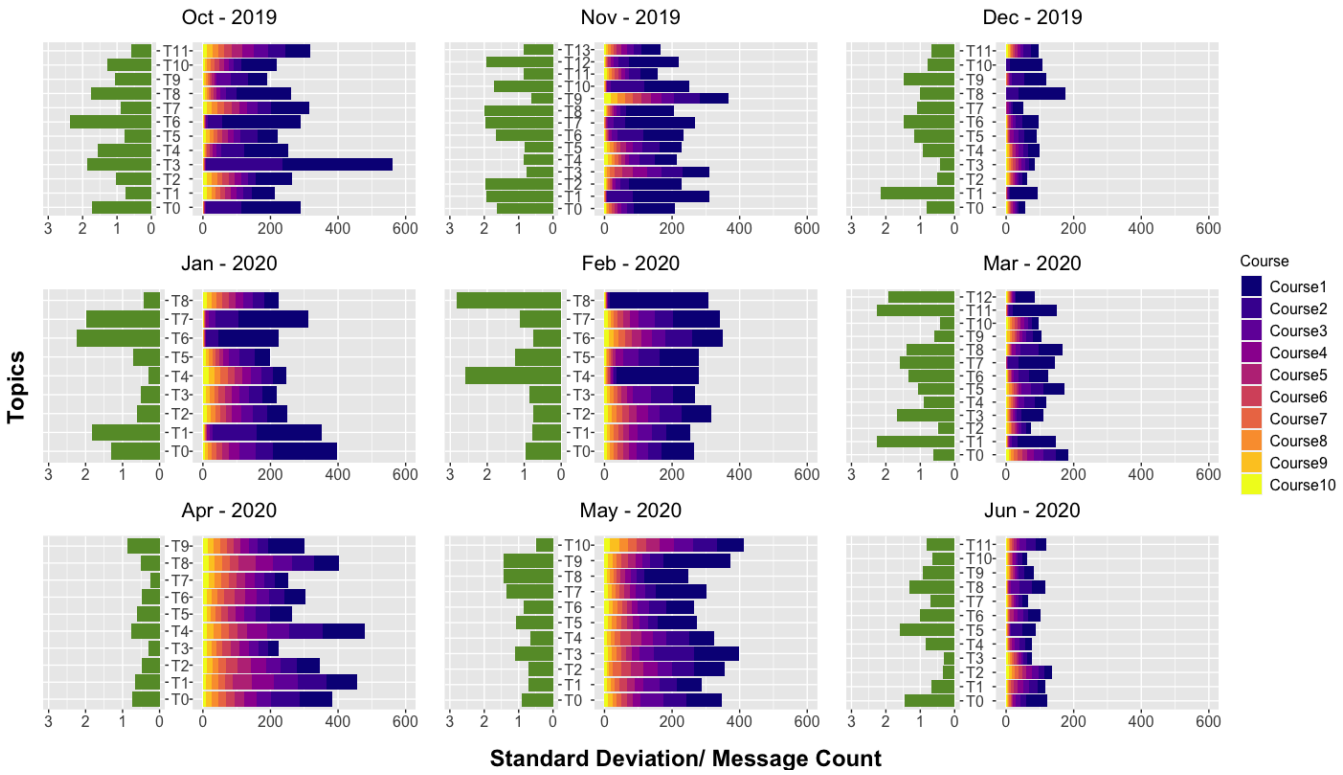

Figure 4: Degree of course-centricity of topics across nine months. Colored gradient bars represent the number of messages in each of the top ten courses that contributed to the topic. Green bars represent the standard deviation of messages from the top ten courses.

6.2 Course-centricity of Topics

In Fig. 4, we demonstrate the course-centricity of topics. Course-centric topics are represented by fewer variations in colors in the gradient bars and a greater standard deviation. By contrast, a more uniform distribution of colors and lower standard deviations such as Socio-cultural Deviance (T2) and Miscellaneous Discourse (T7) in April 2020 demonstrate that a topic is more generally distributed across various courses. We note two key findings. First, we found causal conversations were more generally distributed across courses while topics such as Global Health (T6) and Social Inequities (T7) in January 2020 were more course-centric. Our results suggest a combination of standard deviation and message count can successfully capture variations in topics' course-centricity. Second, we observed a decrease in course-centric topics in the Spring quarter compared to Fall and Winter quarters, which suggests a shift in students' conversations as they transitioned to online learning. To empirically test this shift, we performed a post-hoc Welch's Two Sample t-test to compare the standard deviation of Fall and Winter quarters with that of the Spring quarter. Standard deviations for the two groups differed significantly ($t(5.36) = p < .001$) such that combined, Fall and Winter quarters had a greater standard deviation ($M = .10$) than in Spring quarters ($M = .03$). Notably, the Spring quarter began a few weeks after COVID-19 was declared a pandemic and fully remote learning was implemented. This finding suggests a general change in the online discussion forum landscape, from supporting course-centric content discussion prior to March 2020, to an increased presence of social interactions and discussion on shared live experiences. This might indicate students seeking opportunities to engage in casual interaction that used to take place in the hallways, walking to classes, after classes, or instructors' attempt in facilitating social presence and creating classroom community as well as integrating critical reflections between learning material and contemporary events during remote learning.

6.3 Topic Evolution

Topic chains constructed using WMD illustrate the evolving and emerging topics (Fig. 5). Our findings reveal two topic chains (Chains 12-13) that are consistently present throughout all nine months of the academic year. Some topic chains are limited to a specific quarter (Chains 1-4) which may reflect the classes offered during that quarter. In contrast, other chains notably stop at (Chains 8-10) or start during (Chains 4 and 6) the transition to online learning which may signal the influence of the pandemic on the topics students discussed. We discuss each of these in turn.

Consistently present, Chain 12 began with Public Health-related discussions in October 2019. This topic chain subsequently linked topics of Health Issues, Healthcare & Social Inequities, and Healthcare & Government. In March 2020, this topic evolved into discussions on the Public Health of China. Notably, this shift occurred at the time COVID-19 was declared a pandemic. These themes further transitioned to Public Health Inequities and eventually COVID-19 related discourse in the Spring quarter. Chain 12 also demonstrates that the topics in earlier months were more

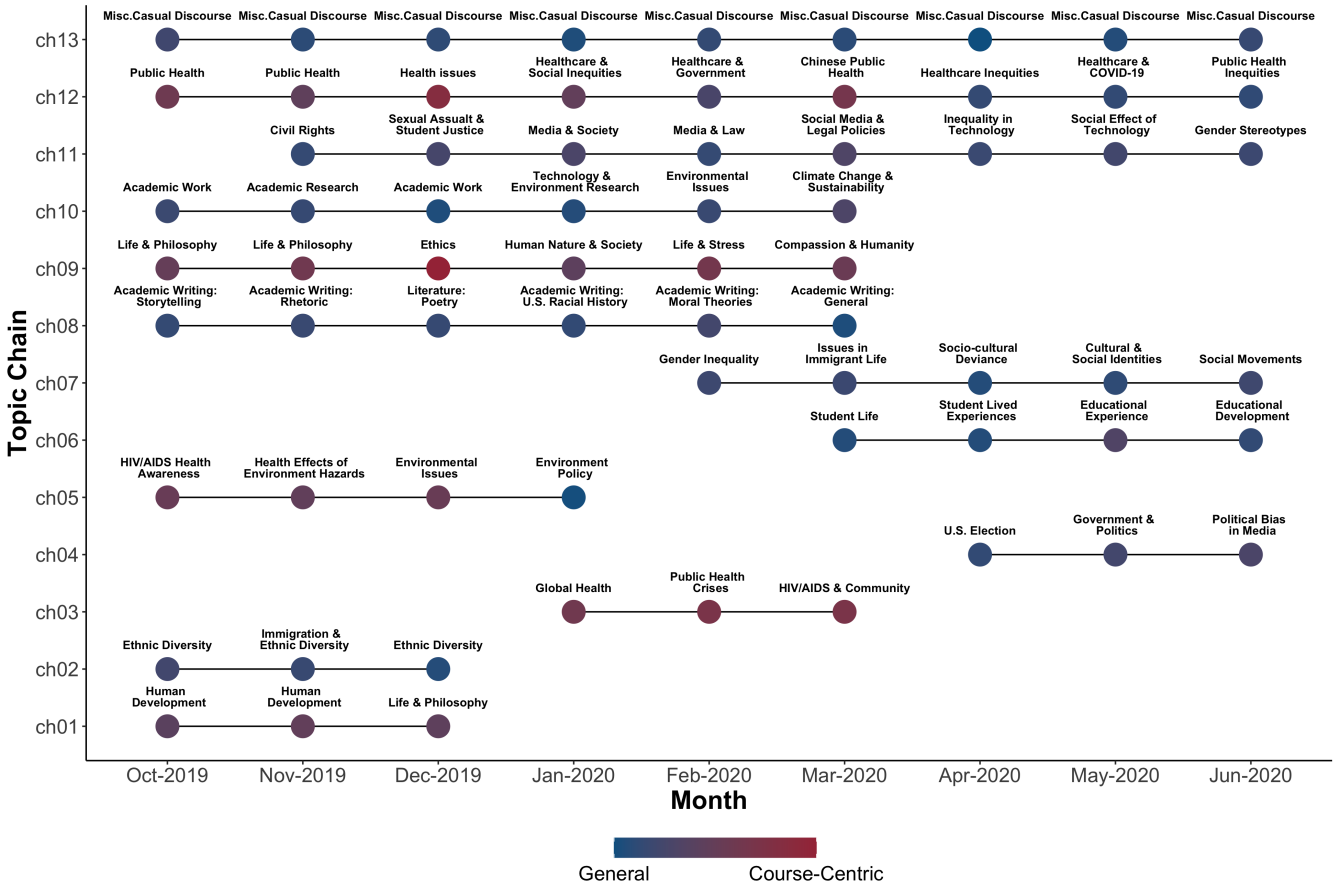

Figure 5: Topic chains identified using the proposed framework. The heat-map scale denotes the course-centricity of a topic.

course-centric, but starting in April 2020 there was a shift towards more general discourse regarding public health inequities and the COVID-19 pandemic in various courses. Another long topic chain (Chain 13) represents miscellaneous messages and casual interactions across all months. This consistent presence implies that online discussions contained some level of student casual interactions prior to and at the onset of the pandemic.

Some topic chains either stopped or began during the onset of the pandemic. For example, Academic writing (Chain 8) and Academic Work (Chain 10) and their respective subtopics were connected from October 2019 to March 2020. These chains highlight the various forms of essay and prompt responses, and reactions to readings or weekly discussion posts that are often found in online discussion classes. Notably, both of these chains contained more "general" topics. We note two possible explanations for the chains' drop-off in March 2020. First, this drop-off may suggest that the courses that prompted these discussions were either no longer available to students or fewer students signed up for these courses during the Spring quarter. Second, it is possible that these discussions surrounding academic work and writing were still taking place, but outside of online discussion forums (e.g., virtual groups, breakout rooms online). Similarly, Chain 9, "Life and Philosophy" related discourse stopped around March 2020. Although most of the posts under these topics were induced by course-related prompts, a closer inspection revealed many posts where students communicated personal experiences and thoughts with their peers. The drop-off for more course-centric chains like Chain 9 could also be due to courses no longer available to students, or perhaps to fewer discussion-based assignments in these courses.

Student life and student's lived experiences were identified as relatively new topics starting in March 2020 (Chain 6). Students' posts included a variety of university-related experiences and major family and life events. A rise in such posts among students during the switch to a fully online education system demonstrated an evolved use of discussion forums to connect with peers and express their thoughts and concerns. These topics were further connected to topics representing shifts in educational experiences. Students expressed their struggles in coping with the sudden transition to online classes and the rapid spread of COVID-19 around the world which added personal or financial difficulties. Other notable posts under this topic included activities that helped them deal with stressful times, their study regime, and new performance evaluation strategies. Other prominent topic chains include Chain 11, "Civil rights" which evolved into related discussions around sexual assault, law and legal policies as-

sociated with social media, and social and gender inequalities. Gender inequality also emerged as a topic in Chain 7 and was linked to other social cause issues regarding immigration, cultural and social movements. Social movements is a diverse topic including terms and events related to feminism, political influence, bias in disseminated information, and fake news.

6.4 Limitations

Our findings demonstrate a clear influence of the COVID-19 pandemic and online learning activities. However, there are a few limitations to our study. The unsupervised nature of topic modeling requires some subjectivity where the authors create topic labels by interpreting the top-ranked words. Future studies should consider blinding the researcher from the month's name when viewing top-ranked words to label topics. Although it is a common practice, the topic labeling could be consequently influenced by the authors' preconceptions of the impact of the pandemic on students' discourse. Moreover, our study does not capture whether a topic is initiated by students or instructors. Although a topic with high course-centricity is influenced by the instructor's prompts to discuss specific academic topics, in some cases, professors also direct and facilitate casual interactions (i.e., introducing themselves, or sharing their experiences during the pandemic). Lastly, our findings reflect discourse changes within asynchronous interactions in online discussions and cannot infer any discourse that took place in synchronous classroom settings or in classes that did not utilize discussion forums for instruction.

6.5 Implications and Future Directions

Large-scale educational data from online discussion forums is an under-excavated gold mine for understanding learning behavior. Educational theories such as constructivism emphasize the role of social interactions and the construction of knowledge through experience and reflection. Analyzing educational discourse can help researchers understand the process of learners' sense-making of new concepts, as well as their attitudes, engagement levels, and experiences.

Extracting meaningful insights from unstructured educational interactions and discussions requires accurate content representation and context consideration. Modeling discourse structure and dynamics pose challenges, such as identifying key topics, tracking the evolution of ideas, and capturing the social and emotional dimensions of communication. To address these challenges, our paper suggests a solution that includes additional LMS features for meaningful interpretation, and effective NLP methods for capturing the dynamic of discourse across different domains and contexts. Specifically, we added LMS features (i.e. message count and standard deviation) to characterize the course-centricity feature of topics to strengthen interpretability. With the case of tracking themes before and after the COVID-19 pandemic, we prove that CTM and WMD can be effective tools to capture emergent topics over time.

This framework is a flexible and adaptable tool that can be adapted to explore other educational research questions in different contexts to investigate teaching and learning behavior in online environments. For instance, future studies could use this method to examine how discourse evolves within specific disciplines (e.g., business courses, STEM courses), potentially for monitoring the consistency of course discourse space. If an instructor wants to understand how students' discourse has changed for a repeatedly offered foundational course, this analytical framework will also be effective for revealing the discourse evolution from past to present to provide insights for the instructor on curriculum design. Future research might also apply it to other formal and informal learning contexts, i.e., MOOCs, social media discussion, and by adding LMS features relevant to self-regulation, which may provide meaningful interpretations of the discourse pattern. While our study only focused on tracking the topics for students who had consistently contributed to the discussion forum, future research may consider how topics evolve across subgroups of students with different demographic backgrounds and individual characteristics such as learner motivation. Although exploring the origins of topics is beyond the scope of this paper, it would be interesting for future research to investigate whether there are differences between conversations initiated by instructors versus students.

7. CONCLUSION

We used an NLP-driven topic detection and tracking approach to detect emergent topics and model the evolution of various topics in students' online academic discourse over time. We demonstrate how this novel NLP technique can be used to provide meaningful insights on large-scale unstructured student data from discussion forums effectively. Our study contributes to the current literature by moving beyond mining static topics from large-scale student discourse to a more process-oriented, temporal lens on how topics emerge, recur and evolve over time. We used contextualized topic models to identify coherent topics from each month, which are more interpretable than traditional models that use only bag-of-words (BoW) features. Some of the identified topics were found to be originating from discussions pertaining to specific courses, while other topics demonstrated the expression of personal opinions and beliefs. Using standard deviation to identify course-centricity, we found that topics posted at the beginning of the pandemic were relatively more general than those before March 2020. This significant increase in casual interactions since Spring 2020 indicates a shift in the discussion forum's function from predominantly enabling academic discourse to increasingly facilitating peer interactions. This evidence supports claims from previous studies that instructors across disciplines were leveraging discussion forums to support social interactions and social learning. The emergent discussion surrounding COVID-19 and other contemporary events in learner discourse suggests that their impacts on students learning and lived experience are not mutually exclusive and are exhibited in both academic and casual discourse. Furthermore, by denoting topics as a set of top representative words based on topic-term probability, we computed the Word Mover's Distance as a semantic similarity metric between adjacent months' topics. The most similar topics were connected and topic chains were constructed to uncover their evolution and identify newer themes. For researchers and practitioners in the EDM community, our proposed approach provides a viable means to characterize topic trends over time in learner discourse at different granularity, such as for specific courses or other online learning contexts. This method might also

be generalized to other types of educational discourse to detect and track specific policy impacts or instructional interventions on students' online discussion activities. Lastly, although this study focused specifically on the events during the 2019-2020 academic year, this approach could be further utilized to understand the temporal dynamics of discussion data in broader contexts and timeframes, i.e., MOOC discussions and social media data.

8. REFERENCES

[1] H. E. Abdelkader, A. G. Gad, A. A. Abohany, and S. E. Sorour. An efficient data mining technique for assessing satisfaction level with online learning for higher education students during the covid-19. *IEEE Access*, 10:6286–6303, 2022.

[2] A. Ahadi, A. Singh, M. Bower, and M. Garrett. Text mining in education—a bibliometrics-based systematic review. *Education Sciences*, 12(3):210, 2022.

[3] R. Alghamdi and K. Alfalqi. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1), 2015.

[4] O. Almatrafi and A. Johri. Systematic review of discussion forums in massive open online courses (moocs). *IEEE Transactions on Learning Technologies*, 12(3):413–428, 2018.

[5] F. Bianchi, S. Terragni, and D. Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021.

[6] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[7] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120. Association for Computing Machinery, 2006.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[9] S. Boon-Itt, Y. Skunkan, et al. Public perception of the covid-19 pandemic on twitter: sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, 6(4):e21978, 2020.

[10] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong. Learning about social learning in moocs: From statistical analysis to generative model. *IEEE transactions on Learning Technologies*, 7(4):346–359, 2014.

[11] D. Bylieva, Z. Bekirogullari, V. Lobatyuk, and T. Nam. Analysis of the consequences of the transition to online learning on the example of mooc philosophy during the covid-19 pandemic. *Humanities & Social Sciences Reviews*, 8(4):1083–1093, 2020.

[12] Z. Cai, B. Eagan, N. Dowell, J. Pennebaker, D. Shaffer, and A. Graesser. Epistemic network analysis and topic modeling for chat data from collaborative learning environment. In *Proceedings of the 10th international conference on educational data mining*, 2017.

[13] C. Chen and Z. Pardos. Applying recent innovations from nlp to mooc student course trajectory modeling. *arXiv preprint arXiv:2001.08333*, 2020.

[14] Y. Chen, B. Yu, X. Zhang, and Y. Yu. Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. In *Proceedings of the 6th International Conference on Learning Analytics and Knowledge*, 2016.

[15] H. Chopra, Y. Lin, M. A. Samadi, J. G. Cavazos, R. Yu, S. Jaquay, and N. Nixon. Modeling student discourse in online discussion forums using semantic similarity based topic chains. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, editors, *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium*, pages 453–457, Cham, 2022. Springer International Publishing.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[17] N. Dowell and V. Kovanovic. Modeling educational discourse with natural language processing. *education*, 64:82, 2022.

[18] N. M. Dowell, C. Brooks, V. Kovanović, S. Joksimović, and D. Gašević. The changing patterns of mooc discourse. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 283–286, 2017.

[19] N. M. Dowell, A. C. Graesser, and Z. Cai. Language and discourse analysis with coh-metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3):72–95, 2016.

[20] N. M. Dowell, Y. Lin, A. Godfrey, and C. Brooks. Exploring the relationship between emergent sociocognitive roles, collaborative problem-solving skills, and outcomes: A group communication analysis. *Journal of Learning Analytics*, 7(1):38–57, 2020.

[21] N. M. Dowell, T. M. Nixon, and A. C. Graesser. Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior research methods*, 51(3):1007–1041, 2019.

[22] N. M. M. Dowell and A. C. Graesser. Modeling learners' cognitive, affective, and social processes through language and discourse. *Journal of Learning Analytics*, 1(3):183–186, 2014.

[23] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth. Unsupervised modeling for understanding mooc discussion forums: a learning analytics approach. In *Proceedings of the fifth international conference on learning analytics and knowledge*, pages 146–150, 2015.

[24] A. Fan, F. Doshi-Velez, and L. Miratrix. Assessing topic model relevance: Evaluation and informative priors. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):210–222, 2019.

[25] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and*

Knowledge Discovery, 9(6):e1332, 2019.

[26] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker, and M. Warschauer. Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44(1):130–160, 2020.

[27] F. B. Goularte, S. M. Nassar, R. Fileto, and H. Saggion. A text summarization method based on fuzzy rules and applicable to automated assessment. *Expert Systems with Applications*, 115:264–275, 2019.

[28] M. Grootendorst. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics., 2020.

[29] 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*, pages 11 – 15, 2008.

[30] L. Hakimi, R. Eynon, and V. A. Murphy. The ethics of using digital trace data in education: A thematic review of the research landscape. *Review of educational research*, 91(5):671–717, 2021.

[31] F. Jian, W. Yajiao, and D. Yuanyuan. Microblog topic evolution computing based on lda algorithm. *Open Physics*, 16(1):509–516, 2018.

[32] S. Joksimović, N. Dowell, O. Poquet, V. Kovanović, D. Gašević, S. Dawson, and A. C. Graesser. Exploring development of social capital in a cmooc through language and discourse. *The Internet and Higher Education*, 36:54–64, 2018.

[33] Z. Kanetaki, C. I. Stergiou, G. Bekas, C. Troussas, and C. Sgouropoulou. Data mining for improving online higher education amidst covid-19 pandemic: A case study in the assessment of engineering students. In *NiDS*, pages 157–165, 2021.

[34] P. Kherwa and P. Bansal. Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24), 2019.

[35] D. Kim and A. H. Oh. Topic chains for understanding a news corpus. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, 2011.

[36] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 2015.

[37] Y. Lin, R. Yu, and N. Dowell. Liwcs the same, not the same: Gendered linguistic signals of performance and experience in online stem courses. In *International Conference on Artificial Intelligence in Education*, pages 333–345. Springer, 2020.

[38] S. Liu, C. Ni, Z. Liu, X. Peng, and H. N. Cheng. Mining individual learning topics in course reviews based on author topic model. *International Journal of Distance Education Technologies (IJDET)*, 15(3):1–14, 2017.

[39] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635, 2020.

[40] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[41] A. Naim. Application of digital technologies for the students with diverse skills during covid: 19. *American Journal of Research in Humanities and Social Sciences*, 1:46–53, 2022.

[42] G. Nanda, K. A. Douglas, D. R. Waller, H. E. Merzdorf, and D. Goldwasser. Analyzing large collections of open-ended feedback from mooc learners using lda topic modeling and qualitative analysis. *IEEE Transactions on Learning Technologies*, 14(2):146–160, 2021.

[43] R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

[44] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, 2010.

[45] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.

[46] M. A. Samadi, J. G. Cavazos, Y. Lin, and N. Nixon. Exploring cultural diversity and collaborative team communication through a dynamical systems lens. 2022.

[47] H. Sha, M. A. Hasan, G. Mohler, and P. J. Brantingham. Dynamic topic modeling of the covid-19 twitter narrative among us governors and cabinet executives. *arXiv preprint arXiv:2004.11692*, 2020.

[48] S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker, and D. Gasevic. Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, 42(1):85–106, 2017.

[49] A. Srivastava and C. Sutton. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[50] W. Strielkowski. Covid-19 pandemic and the digital revolution in academia and higher education. 2020.

[51] C.-H. Tu. On-line learning migration: From social learning theory to social presence theory in a cmc environment. *Journal of network and computer applications*, 23(1):27–37, 2000.

[52] L. S. Vygotsky and M. Cole. *Mind in society: Development of higher psychological processes*. Harvard university press, 1978.

[53] C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, page 579–586. AUAI Press, 2008.

[54] X. Wang, D. Yang, M. Wen, K. Koedinger, and C. P. Rosé. Investigating how student's cognitive behavior in mooc discussion forums affect learning gains. *International Educational Data Mining Society*, 2015.

[55] M. Wen, D. Yang, and C. Rose. Sentiment analysis in mooc discussion forums: What does it tell us? In

Educational data mining 2014. Citeseer, 2014.

[56] A. F. Wise, Y. Cui, and J. Vytasek. Bringing order to chaos in mooc discussion forums with content-related thread identification. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 188–197, 2016.

[57] World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020., 2020. https://www.who.int.

[58] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du, and W. Buntine. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720, 2021.