Modeling Student Discourse in Online Discussion Forums Using Semantic Similarity Based Topic Chains

Harshita Chopra[1(✉)], Yiwen Lin[2], Mohammad Amin Samadi[2], Jacqueline Guadalupe Cavazos[2], Renzhe Yu[2], Spencer Jaquay[2], and Nia Nixon[2]

[1] GGS Indraprastha University, Delhi, India
harshitachopra3@gmail.com
[2] University of California, Irvine, USA

Abstract. Students' conversations in academic settings evolve over time and can be affected by events such as the COVID-19 pandemic. In this paper, we employ a Contextualized Topic Modeling technique to detect coherent topics from students' posts in online discussion forums. We construct topic chains by connecting semantically similar topics across months using Word Mover's Distance. Consistent academic discourse and contemporary events such as the COVID-19 outbreak and the Black Lives Matter movement were found among prominent topics. In later months, new themes around students' lived experiences emerged and evolved into discussions reflecting the shift in educational experiences. Results revealed a significant increase in more general topics after the onset of pandemic. Our proposed framework can also be applied to other contexts investigating temporal topic trends in large-scale text data.

Keywords: Text mining · Discourse analysis · Topic modeling

1 Introduction

The onset of the COVID-19 pandemic prompted an urgent shift to online education and created a nontrivial disruption in students' educational experience that affected their academic engagement and mental health [5]. The rapidly changing nature of the pandemic underscores the need for an automated way of detecting the temporal dynamics of themes discussed online and the potential insights they give on its influence on education. Here, we aim to leverage Natural Language Processing (NLP) techniques to capture emergent topics and temporal evolution of undergraduates' online discourse in discussion forums in the months prior to and throughout the pandemic. We employed the Combined Topic Model (CombinedTM) [1] to extract coherent themes that emerged monthly and used Word Mover's Distance (WMD) to construct topic chains by computing the semantic similarity between topics across adjacent months. Additionally, we propose

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13356, pp. 453–457, 2022.
https://doi.org/10.1007/978-3-031-11647-6_91

a measure of course-centricity to distinguish topics that are more specific to certain courses from those which represent broader themes that were observed across multiple courses.

2 Background

Topic modeling methods such as Latent Dirichlet Allocation [2] have been used to extract static themes in learner-generated data and to study the impact of the pandemic on teaching and learning in higher education [8]. However, most of these studies have shown limited capacity to detect coherent topics and do not reflect temporal changes in themes discussed online. Given the rapid changes brought to educational settings, we seek to examine how topics emerge, recur and evolve in student discourse.

Recent advances in deep learning have introduced the combination of neural networks and transformer-based techniques to yield topics that are more coherent and interpretable than traditional models. In this study, we used CombinedTM, a recently proposed neural topic model that uses a Bag of Words (BoW) document representation concatenated with the contextualized document representation from Sentence-BERT [7].

To connect different topics temporally, previous studies have used traditional similarity metrics [3]. By contrast, we used WMD [4] to track topics that represent a similar broad theme but depict a change in context over time. WMD measures the dissimilarity between two text documents, leveraging the power of word embeddings [6], even if they do not have any words in common. By exploring the temporal characteristics of learner discourse during this critical time, we aim to enhance our understanding of the influence of the pandemic and policy responses on learning activities.

3 Data and Methodology

The dataset was obtained from the online discussion forums on the learning management system at a large public university in the United States during the academic year from October 2019 to June 2020. We retained posts generated from the same individuals across months, and removed posts that contained less than two words or five characters. A total of 32,409 posts created by 449 students across 636 courses were retrieved and preprocessed to retain relevant tokens.

We trained CombinedTM on the discussion posts for each month separately. The BoW vocabulary was constructed by retrieving the top 10,000 words with maximum Term Frequency - Inverse Document Frequency weights and Sentence-BERT was used to obtain encodings of the posts. To determine the optimal number of topics (K), we ran the models for each month with K ranging from 5 to 15 topics and evaluated them on the three metrics used by [1]. To determine the degree of course-centricity, we examined how each topic was distributed across courses. We assigned each post a topic with the highest probability. For each topic, the frequencies of the posts for the top-$N(= 10)$ most common courses

were used to calculate the standard deviation (σ). A lower value of σ denoted a relatively uniform distribution of courses in a topic, suggesting a topic represents a broader theme that is more generally distributed across multiple courses. A higher value of σ denoted a skewed distribution where very few specific courses dominate the discussion, showing that the topic is more "course-centric".

We used WMD to measure the semantic or contextual similarity between every pair of topics in adjacent months. A Word2Vec model [6] was trained on the entire corpus to obtain 100-dimensional word embeddings. Considering each topic as a list of top-30 representative words, we computed the WMD between all topic pairs belonging to adjacent months (m_t and m_{t+1}). For every topic in m_t, we selected the topic having the least WMD (the most similar) from m_{t+1}. To avoid multiple topics in m_t getting mapped to the same topic in month m_{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 and found all simple paths from each root to leaf. These directed paths are referred to as "topic chains".

4 Results and Discussion

The topic modeling resulted in 8–13 number of optimal topics per month, including students' lived experiences and contemporary events such as social justice movements, which demonstrate sociocultural influences on learning. Details on the topics and top-ranked words are made publicly available[1].

We empirically tested a shift in course centricity with a post-hoc Welch's Two Sample t-test to compare the degree of variability in Fall 2019 and Winter 2020 with that of Spring 2020. Fall and Winter quarters had a greater standard deviation ($M = .10$) than in the Spring quarter ($M = .03$), $t(5.36) = p < .001$. This finding shows that topics became less course-specific in the Spring, which began a few weeks after fully remote learning was implemented due to the COVID-19 pandemic, than in the previous two quarters. Although online forums mainly serve as a place for course-oriented discussions, the emergence of more general topics indicates a common or shared online experience across different courses.

Amongst the identified topic chains (Fig. 1), the top two most consistent themes were casual interactions (Chain 13) and Public Health-related discussion (Chain 12). Chain 12 demonstrated that discourse around public health began as course-centric topics in earlier months and later became more general regarding pandemic-related health inequities. This suggests that public health discussions expanded beyond corresponding courses, became a shared concern and arose in broader student discourse during the pandemic.

Student Life emerged as a relatively new topic starting Mar-2020 (Chain 6). Students' posts included university-related experiences, and major family and life events. A rise in such posts demonstrated an evolved use of online discussion forums to connect with peers during remote learning. This information suggests

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

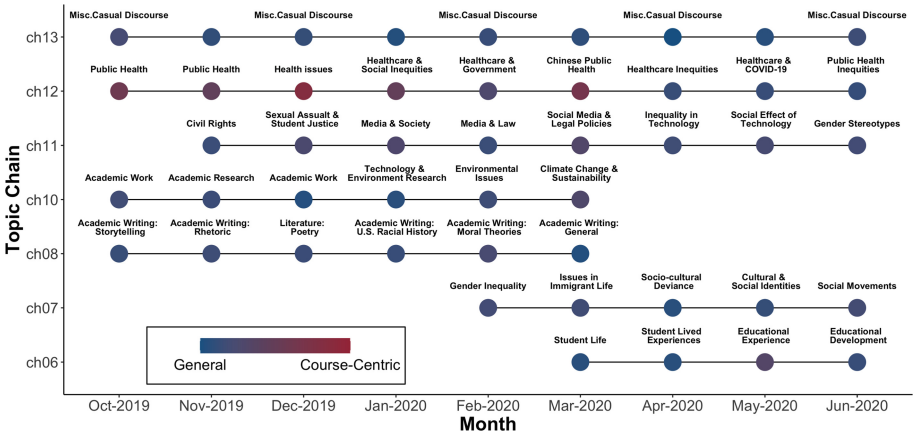

Fig. 1. Topic chains colored by course-centricity of each topic.

the shift in discussion forum's role in providing space for academic discussion to sharing experiences and building social connections in the classroom community. Future studies are needed to investigate how this change might influence learners' sense of belonging during remote learning.

Conclusion. Our study contributes to the literature by moving beyond mining static topics from large-scale discussion forums, towards a more process-oriented, temporal technique of modeling topics. For researchers and practitioners in the AIED community, our proposed approach provides a viable means to analyze the development of discourse in online educational environments in response to certain events or introduction of new policies.

References

1. Bianchi, F., Terragni, S., Hovy, D.: 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, vol. 2 (2021)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Kim, D., Oh, A.H.: 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)
4. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML 2015 (2015)
5. Means, B., Neisler, J., et al.: Suddenly online: a national survey of undergraduates during the Covid-19 pandemic. Technical report, Digital Promise (2020)
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, vol. 26. Curran Associates, Inc. (2013)

7. Reimers, N., Gurevych, I.: 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)
8. Vijayan, R.: Teaching and learning during the COVID-19 pandemic: a topic modeling study. Educ. Sci. **11**(7), 347 (2021)