Exploring the Implementation of NLP Topic Modeling for Understanding the Dynamics of Informal Learning in an AI Painting Community

Ran Bi*, Shiyao Wei*
SAS Institute, Florida State University
ran.bi@sas.com,
sw22b@fsu.edu

ABSTRACT
Informal learning is a significant part of lifelong learning. The rise of online communities as a new venue for informal learning has led to an increase in the availability of discourse data. As the dataset grows, it is feasible for scholars to understand the learning dynamics of these communities. However, the manual coding and analysis of such large datasets can be cost-prohibitive. Natural Language Processing (NLP) has been demonstrated to be a viable solution for analyzing large datasets in educational contexts. In this paper, we explore the application of NLP topic modeling method, Latent Dirichlet allocation (LDA), in understanding informal learning dynamic within an AI painting community. We collected data in two months from November 7, 2022, to January 8, 2023, and our findings show that major topics discussed in the space are around ethics, models, and procedures of AI painting, and topics updated over two months.

Keywords
Topic modeling, Affinity Space, LDA, AI Painting, Informal Learning

1. INTRODUCTION
The first and second decades of the 21st century have seen the emergence of online communities, such as subreddits on Reddit and groups on Facebook. These communities provide a platform for interest-driven learning outside of formal education settings [11]. Studies have shown that online spaces, particularly those in remote areas, provide individuals with a shared space to learn diverse knowledge [4], such as literacy [3] and disease management [15]. Additionally, online affinity spaces bridge the gap between socioeconomic, ethnic, and social groups, allowing learners to communicate freely around topics of interest [4].

As the Internet becomes more prevalent, there is a growing amount of data available for studying online affinity spaces [8]. However, as the size of data increases, so does the cost of hand-coding it. Traditional qualitative coding requires researchers to read and understand thousands of data points [12], which can be costly and time-consuming. To address this issue, researchers have been exploring alternative methods, such as natural language processing (NLP), to get a snapshot of the data before embarking on the hand-

R. Bi and S. Wei. Exploring the implementation of nlp topic modeling for understanding the dynamics of informal learning in an ai painting community. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 434–437, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115754

AI painting is a new application of generative AI. It works by using algorithms to analyze and learn from images available on the Internet and input specified by humans [13]. The algorithm generates new images in adherence to the aesthetics it has learned. AI painting has attracted public attention and sparked many discussions in online communities due to its potential and associated risks. For example, the subreddit r/StableDiffusion is a prevalent community where participants gather, share, ask, and debate around AI painting issues. In previous work, we hand-coded 2,291 posts and comments in r/StableDiffusion and found eight major topics of discussion: algorithm & model, application, data, entertaining, ethics & social implications, hardware, off-topic, and procedure. In this paper, we use Latent Dirichlet allocation to identify the major topics discussed in the space and determine if there are changes within 2 months.

2. RELATED WORK
2.1 Affinity Space and Informal Learning
Online affinity spaces have garnered the attention of researchers in education field. Affinity spaces are a form of public pedagogy in informal learning [6]. In these spaces, learners exchange information about shared passions through design and resources [6, 15]. Affinity spaces help learners prepare for their lifelong learning journey outside of traditional educational environments.

Discussions play a crucial role in the information-exchange process within affinity spaces. Understanding the dynamics of these discussions is essential for studying informal learning. Recent studies on the discussion patterns of affinity spaces [14, 15] have identified key content types on online social network sites and different behaviors between key and other actors. While [14] collected 514 posts discussing disease management, they did not examine the change of topics over time. Additionally, previous studies have not addressed the issue of how the topics in a technology-focused affinity space change. Through this study, we aim to utilize topic modeling to analyze the larger scale of discourse data and identify the dynamics of the space.

2.2 Topic Modeling in Discourse Analysis
Topic modeling is an NLP method applied in discourse analysis. One of the most widely used topic modeling methods is Latent Dirichlet Allocation (LDA), which derives probabilities of words belonging to topics (clusters of semantically related words) from textual data [21]. Another study [20] investigated the potential of using LDA to explore topics emerged in social media data during the COVID-19 pandemic and found that LDA is useful for stakeholders to understand the most discussed topics in the field.

The gap in literature identified frames our research question and the methods discussed above contribute to the choice of our research methods. Our research questions are as follows:

RQ1: *What are the topics most discussed in the subreddit in 2 months?*

RQ2: *How do topics change in an AI painting affinity space in 2 months?*

3. METHOD

3.1 Data Collection and Cleaning

In this research, we aim to observe the topics discussed and how topic changed in an AI painting affinity space. Thus, we chose one highly frequented platform for discussion surrounding AI-generated painting as the observation site. We observed the community for 2 months, utilizing the Pushshift API [1]. All posts and comments were obtained within a specific time frame, from November 7th, 2022, to January 8th, 2023. The sample included 14,319 posts and 172,770 comments. The 2-month period included workdays, weekends, and vacation season to help us better understand the trend and dynamics of informal learning in the space. In terms of data cleaning, a flexible approach was implemented. Firstly, comments were removed if they were not associated with posts within the designated time frame. Secondly, all posts were concatenated based on their title, text, and comments, being treated as a single paragraph.

3.2 Data Analyzing and Visualization

In the process of analyzing data, a time series analysis was conducted on the number of posts, comments, and subscribers of the Stable Diffusion channel over a period of two months (9 weeks). It indicates an increase in subscribers from 80,000 to 116,000 at a steady rate. Furthermore, the time series plot of comments revealed three peaks during the two-month period. The first peak occurred during the week of Thanksgiving, due to the viral spread of AI-generated holiday greeting graphs. The other two peaks occurred two weeks prior to Christmas. Based on our analysis, we noticed a seasonality of increased comments on weekends as opposed to a higher frequency of post submissions on weekdays. Additionally, text mining and topic modeling using LDA was conducted, yielding interesting results that warrant further investigation. Data visualization is achieved by using LDAvis [16], an interactive visualization of topics estimated using LDA. As shown in Figure 1, bubble graph refers to different topics emerged from the material, with a red bubble highlighted. The bar chart refers to the frequency of top 30 terms related to topic 1 that appeared in the context of topic 1. The slide bar could adjust the relevance parameter of terms. The numbers of week mentioned in this paper represent the order of week in a year, for example, week 45 is the 45th week in 2022.

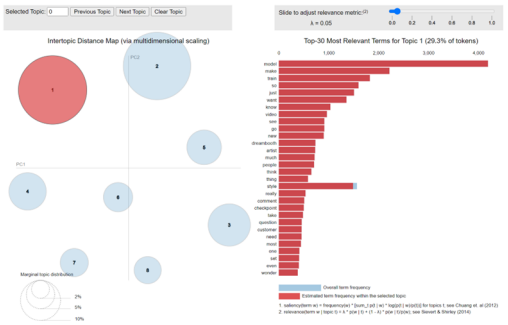

Figure 1. Topic modeling of posts-only data

4. RESULTS

4.1 What are the Topics Most Discussed in the Subreddit in 2 Months?

When we examined the results of posts only, topics related to models, such as training the model and Dreambooth, are the most discussed. Ethical and social implications, including keywords such as art and artists, are also mentioned in the first category. However, when analyzing both posts and comments, the topics become clearer, as shown in Figures 1 and 2.

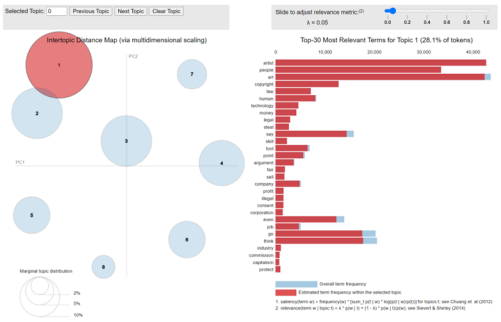

Figure 2. Topic modeling of posts and comments data

In the results of posts-and-comments, topics related to ethics, such as artists and art, are separated from the previous first category. We found that the results of LDA partly align with the hand-coding results from our previous research. In the results of all posts and comments, the second most discussed topic is about models, which contains words such as "model", "train", and "training". In topic 3, words related to procedures, such as "run", "file", "folder", and "download" cluster together. In topic 5, words related to applications, such as "video", "game", and "life" emerged. Topic 6 is about all prompts used in the process of generating images, such as "prompt", "picture", "text", "girl", and "man".

4.2 How did Topics Change in an AI painting Affinity Space in 2 Months?

In our observation of weekly differences, we found that in Week 45, the discussion of ethics in Topic 2 focused on the issue of copyright, as shown in Figure 3. By adjusting the relevance metric to 0, we observed an increase in the weight of the keyword "copyright".

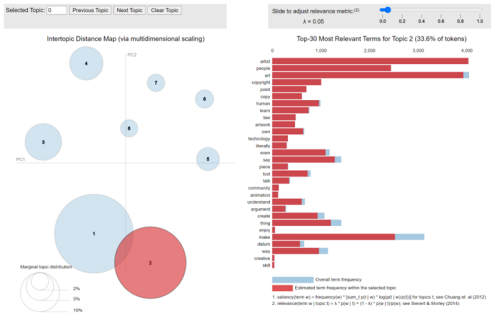

Figure 3. Topic modeling result from week 45

In Week 46, as shown in Figure 4, participants were more concerned about job and industry in the context of ethics. Additionally, compared to the previous week, there was a new discussion on watermarks, due to the release of Stable Diffusion 2 (SD2) which tends to generate images with watermarks, which many users were complaining about.

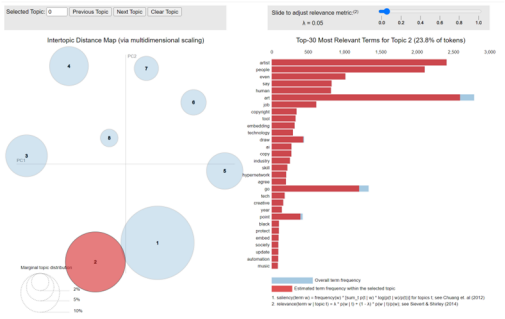

Figure 4. Topic modeling result from week 46

In Week 47, as shown in Figure 5, we noticed a different keyword, "censorship." Upon further examination of the original data, we found that it was related to the release of SD2.

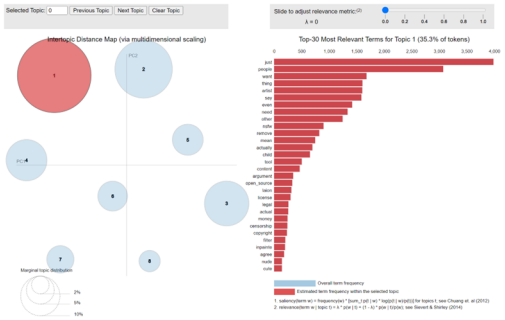

Figure 5. Topic modeling result from week 47

5. DISCUSSIONS

In alignment with previous research [5, 18], in this paper, we believe that LDA can only provide a glimpse of the data and capture certain keywords in the discussion. To gain more in-depth insights, researchers must go back to the data and explore the reasons behind why these keywords appear. However, LDA can save time by reducing the need to read less important data and improve the efficiency of analyzing discourse data in social media. Affinity space as an information hub, dispersed knowledge pattern, time-sensitivity of topics, and limitations and future research are discussed below.

5.1 Information Hub for Interested Learners

Affinity spaces serve as an information hub for all interested learners, and our results show that they contain mainstream topics related to AI painting, aligning with our previous hand-coding results. Topics related to ethics and social implications, such as art, artists, industry, jobs, and copyright, consistently took the first or second position throughout 2-month data. Similarly, topics related to applications consistently appeared throughout 2 months, although with a lower ranking. Procedures and algorithm & models were the second most discussed topics. Also, users participated in much discussion about prompts and data in the subreddit, echoing the open-source tradition in coding community [7]. Discussions about hardware appeared in several weeks and some keywords related to hardware were mixed in the algorithm and models category. Some entertaining content was also mixed in off-task categories.

According to [4], the subreddit r/StableDiffusion acts as an affinity space where learners can share their experiences and knowledge surrounding AI-generative painting, specifically about Stable Diffusion. This generator allows individuals to gather and explore sets of signs and potential relationships among signs [4]. Unlike other learning environments such as bootcamps that cater to a specific level of skill, r/StableDiffusion welcomes both beginners and experts alike [4]. The community encourages both extensive and intensive learning [4], with members frequently sharing analyses of results, algorithms, and models, providing abundant resources for novice learners entering the space.

5.2 Dispersed Knowledge of Affinity Space

During our topic modeling analysis, we observed that learners engage in sharing behaviors that connect to other sites, such as “png” and “github”, which are external to the subreddit. This reveals that space enables users to actively participate in sharing and learning beyond its confines. The distributed nature of knowledge in network-like formats means that there are no strict boundaries or limitations on what learners can access, which encourages their agency [19]. Freedom and choice are vital components of informal learning [17], and the dispersed nature of the resources connected by a network-like format enables learners to select what interests them the most, and continue their learning journey accordingly.

5.3 Time-sensitiveness of Affinity Space

We found that the affinity space is time-sensitive, meaning that users promptly respond to updates of Stable Diffusion and other related news in AI painting. For example, even before the public release of Stable Diffusion 2 on November 24, 2022, users were discussing the watermark issue in SD2 in Week 46 (November 7-13, 2022). Similarly, the launch of ChatGPT on November 30, 2022 was also discussed in the following week's discussion. This finding adds to current understanding of affinity space and informal learning, especially a few principles mentioned in [4]. Learners in affinity space proactively react to the development and updates of the software and might change the development of the software itself.

Time-sensitiveness matters because it contributes to the learning material development in the affinity space. Affinity space is a public pedagogy [6]. People come here for up-to-date experience and knowledge, thus, the immediacy of sharing and response to the software in the space are useful learning materials for all learners in the space.

5.4 Limitations and Further Research

While our analysis captured the topics that emerged during the two-month period of our data collection, we acknowledge that the dynamics of the community are constantly evolving. In addition, comparing the keywords from each week proved challenging due to the sheer volume of data. Future research could benefit from using dynamic topic modeling [2], another NLP methods in discourse analysis, to achieve a more in-depth understanding of how the community's discourse and topics of discussion evolve over time.

6. CONCLUSION

In this paper, we report on the progress of using Latent Dirichlet Allocation (LDA) to capture the dynamics of topics in an AI painting affinity space. We collected data over a two-month period, from November 7, 2022, to January 8, 2023. Our findings indicate that the community's primary topics revolve around ethics, models, and procedures, and that these topics evolved over the course of the two-month period.

7. REFERENCES

[1] Baumgartner, J. et al. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*. 14, (May 2020), 830–839. DOI:https://doi.org/10.1609/icwsm.v14i1.7347.

[2] Blei, D.M. and Lafferty, J.D. 2006. Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning - ICML '06* (New York, New York, USA, 2006), 113–120.

[3] C. Lammers, J. et al. 2012. Toward an affinity space methodology: Considerations for literacy research. *English Teaching*. 11, 2 (Jul. 2012), 44-n/a.

[4] Gee, J.P. 2005. Semiotic social spaces and affinity spaces: from *The Age of Mythology* to today's schools. *Beyond Communities of Practice*. Cambridge University Press. 214–232.

[5] Gencoglu, B. et al. 2023. Machine and expert judgments of student perceptions of teaching behavior in secondary education: Added value of topic modeling with big data. *Computers & Education*. 193, (Feb. 2023), 104682. DOI:https://doi.org/10.1016/j.compedu.2022.104682.

[6] Hayes, E.R. and Gee, J.P. 2010. *Handbook of Public Pedagogy*. Routledge.

[7] Heller, B. et al. 2011. Visualizing collaboration and influence in the open-source software community. *Proceedings of the 8th Working Conference on Mining Software Repositories* (New York, NY, USA, May 2011), 223–226.

[8] Hewson, C. 2020. Qualitative Approaches in Internet-Mediated Research: Opportunities, Issues, Possibilities. *The Oxford Handbook of Qualitative Research*. Oxford University Press. 633–673.

[9] Jacobs, T. and Tschötschel, R. 2019. Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*. 22, 5 (Sep. 2019), 469–485. DOI:https://doi.org/10.1080/13645579.2019.1576317.

[10] Nelson, L.K. et al. 2021. The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods. *Sociological Methods & Research*. 50, 1 (Feb. 2021), 202–237. DOI:https://doi.org/10.1177/0049124118769114.

[11] Peters, M. and Romero, M. 2019. Lifelong learning ecologies in online higher education: Students' engagement in the continuum between formal and informal learning. *British Journal of Educational Technology*. 50, 4 (Jul. 2019), 1729–1743. DOI:https://doi.org/10.1111/bjet.12803.

[12] Prior, L. 2020. Content Analysis. *The Oxford Handbook of Qualitative Research*. P. Leavy, ed. Oxford University Press.

[13] Reed, S. et al. 2016. Generative adversarial text to image synthesis. (2016), 1060–1069.

[14] Sharma, P. et al. 2021. Knowledge sharing discourse types used by key actors in online affinity spaces. *Information and Learning Sciences*. 122, 9/10 (Sep. 2021), 671–687. DOI:https://doi.org/10.1108/ILS-09-2020-0211.

[15] Sharma, P. and Land, S. 2019. Patterns of knowledge sharing in an online affinity space for diabetes. *Educational Technology Research and Development*. 67, 2 (Apr. 2019), 247–275. DOI:https://doi.org/10.1007/s11423-018-9609-7.

[16] Sievert, C. and Shirley, K. 2014. LDAvis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (2014), 63–70.

[17] Song, D. and Bonk, C.J. 2016. Motivational factors in self-directed informal learning from online learning resources. *Cogent Education*. 3, 1 (Dec. 2016), 1205838. DOI:https://doi.org/10.1080/2331186X.2016.1205838.

[18] Törnberg, A. and Törnberg, P. 2016. Muslims in social media discourse: Combining topic modeling and critical discourse analysis. *Discourse, Context & Media*. 13, (Sep. 2016), 132–142. DOI:https://doi.org/10.1016/j.dcm.2016.04.003.

[19] Wu, G.C.-H. and Chao, Y.-C.J. 2015. Learners' agency in a Facebook-mediated community. *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (Dec. 2015), 558–563.

[20] Xue, J. et al. 2020. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. *PLOS ONE*. 15, 9 (Sep. 2020), e0239441. DOI:https://doi.org/10.1371/journal.pone.0239441.

[21] Zamani, M. et al. 2020. Understanding Weekly COVID-19 Concerns through Dynamic Content-Specific LDA Topic Modeling. *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science* (Stroudsburg, PA, USA, 2020), 193–198.