International Society of the Learning Sciences

Exploring Impacts of Arts Participation through Automated Inductive Annotation

Seth Corrigan, University of California-Irvine, scorrig1@uci.edu
Santiago Ojeda-Ramirez, University of California-Irvine, sojedara@uci.edu
Maggie Dahn, University of California-Irvine, dahnm@uci.edu
Daniela DiGiacomo, University of Kentucky, daniela.digiacomo@uky.edu
Julian Sefton-Green, Deakin University, julian.seftongreen@deakin.edu.au
Kylie Peppler, University of California-Irvine, kpeppler@uci.edu

Abstract: Researchers have worked to identify the varied outcomes associated with participation in the arts. Historically, those efforts have tended to take a 'top down' approach in that they focus on specific types of outcomes, or existing theoretical orientations. In this paper, we demonstrate an approach to annotating transcribed interviews that blends inductive human coding with machine learning in order to support 'bottom-up' discovery of previously unrecognized outcomes. We utilize a recent collection of retrospective semi-structured interviews with 102 international participants in the arts. The annotation approach yields a system for clustering and cataloging the interviews according to a set of modeled topics that facilitates corpus-based research, and may promote identification of new categories of outcomes. The work is one part of a project to create a community-driven taxonomy of outcomes of arts participation that encompass a universe of reported outcomes not delimited by domain, type, or theoretical orientations of stakeholders.

Introduction

This paper presents a method for automating annotation of large corpora as a support to qualitative research efforts. The method is demonstrated in the context of an ongoing qualitative research effort to understand outcomes associated with arts participation. One of the benefits of the approach is found in its connection to grounded theory– permitting researchers to work in an inductive manner to identify outcomes that may not be prevalent in current discourses.

The method for automating annotation of large corpora has relevance for researchers investigating outcomes in a broad range of domains. Here, we apply the approach to identify outcomes associated with arts participation that may not be available or prevalent in current discussions. Learning scientists and arts education researchers have carried out extensive efforts to identify and document learning processes and associated benefits associated with arts participation (e.g., Halverson, 2013; Halverson & Sawyer, 2022). But at least two factors influence the ability of many of these studies to discover the full universe of outcomes that participants experience.

First, where studies take a deductive, or 'top-down' approach and investigate outcomes of a single type, they necessarily delimit the types of outcomes they can expect to find. Decisions to narrow such a search are often practical. But they also have epistemic impacts, as they may limit discovery of a broader variety of outcomes. This scenario potentially leads to a trade-off between expediency and breadth of the given investigation (Wallace, 2007). Inductive, and mixed inductive-deductive methods can be leveraged to identify potentially new categories within data, without relying on pre-existing theories (Wallace, 2007; Bingham and Witowsky, 2021).

Second, where program location, participant characteristics and other factors impact the types of outcomes participants experience and report, research with limited or homogenous samples will be less likely to surface a wide variety of outcomes. By increasing the size and variability of their participant population, researchers can increase the chances they will identify a larger set of unique outcomes. But in the case of interview data, working with large corpora presents its own problems - as the sizes of interview corpora grow for example, they become increasingly difficult to organize and manage in ways that support easy browsing, querying and filtering for further research (Ingwersen, 1992; Uren et al., 2006). This is a recognized issue in the fields of corpus linguistics and information science more generally, and it has led to well developed *annotation procedures* for indexing corpora for utility and ease-of-use by researchers and other stakeholders (Wallis, 2007; Gries & Berez, 2017).

Annotation, sometimes referred to as 'tagging', can be thought of as a process of labeling specific rows or subsets of a corpus with additional information such as linguistic, pragmatic or semantic markers in order to make the corpus more useful for a given goal (Hovy & Lavid, 2010). Such annotations can be made by hand or automated through use of machine learning or other machine-based approaches. Blended annotation procedures can be thought of as using both human annotations of the data as well as automated annotations. Given the twofold problem surfaced here, we expect there is benefit to a blended inductive approach to corpus annotation that can simultaneously support efforts to discover a more complete universe of outcomes associated with arts practice while also facilitating use of the corpus.

Purpose

The purpose of the current study is to demonstrate a blended approach to corpus annotation that leverages inductive human coding and unsupervised machine learning to support discovery of potentially novel, or previously unrecognized classes of outcomes reported by participants in the arts. Use of this inductive blended annotation approach leads to clustering and labeling an extensive set of outcomes described by adult interviewees who participated in the arts or educational arts programs as children or youth. We refer to the approach as inductive because it utilizes patterns in participants' speech to identify outcomes, as opposed to beginning with a more narrow set of established categories or theories of outcomes. It is 'blended' because it leverages both human-generated annotations as well as annotations derived from machine learning methods.

This inductive blended approach has at least two virtues. First, it supports inquiry into alternative ways of clustering participants' outcome statements and potential meanings associated with the resulting clusters, without engaging with existing work or theories that may narrow or limit that exploration. As a result, the approach allows for discovery of a broader universe of outcomes of arts participation. Second, the annotation approach also provides the means to catalog, or index the corpus with reference to types of outcome statements contained in each interview and which utterances contain them. When applied in the context of qualitative research conducted on a corpus, the blended annotation approach yields cluster labels and accompanying sets of candidate topics that can be used to index and organize the corpus in order to ease browsing, searching, and filtering content for further research.

Background

The conviction that participation in the arts has impacts that range beyond their aesthetic value has a history dating at least as far back as Aristotle (Belfiore & Bennett, 2007). Since 2015, efforts have grown to better understand those impacts (Sol, Gustren, Nelhaus et al., 2021). To name a few, these investigations include the benefits of arts participation for academic achievement (Guhn et al., 2020; Jindal-Snape, 2018), student engagement (Walker et al., 2011), mental health and wellness (Kosma et al., 2020; Stuckey & Nobel, 2010), executive functioning (Holochwost et al., 2017), confidence (Simpson Steele, 2019), social relationships (Dadswell et al., 2020), community building and connection (Catterall, 2009; Catterall et al., 2012; Stevenson & Deasy, 2005), and occupational outcomes (Betts, 2006), among others. Less prevalent are studies such as Matarasso (1997), and Merli, (2003) who have taken a more general approach, aiming to identify outcomes across a larger number of dimensions, though even they stay within the single general category of "social impacts."

Benefit of bottom-up annotation for discovery

In addition to such focused, deductive approaches to investigating impacts of arts participation, the field may benefit from more open and inductive investigations as well. Inductive, or 'bottom-up' approaches to qualitative research include a range of processes for reading and interpreting text to then develop concepts, themes or models to aid subsequent interpretation of that same data (Boyatzis, 1998; Corbin & Strauss, 1990). In that sense, inductive methods may be thought of as being "data first," relying on forms of inductive reasoning to surface important themes, topics, and models from the data itself without preconceived notions or limitations on the number or types of outcomes to be found. As a result, use of inductive methods can support exploratory efforts to identify new ways of categorizing, and therefore understanding the data, and in the current case, help to explore potentially novel or less well-known outcomes of arts participation.

Corpus annotation to facilitate exploration and research in the learning sciences

Efforts to identify and explore a more complete universe of outcomes from arts participation, must consider the characteristics of the participant pool. Examining interview data from a large group of diverse participants can support discovery of a more complete universe of outcomes. But when interview data is being used, large amounts of text data can result, and large corpora present problems related to data management and information retrieval (Ingwersen, 1992). In particular, many of the problems associated with large corpora relate to organizing the data so that it can be easily browsed, sorted and filtered, making it more useful for researchers and other stakeholders.

This is a standing problem in the fields of corpus linguistics (Leech, 2005), information science and knowledge management (Uren, Cimiano, Iria, et al. 2006), that has been at least partially solved through use of annotation processes (Wallis, 2007). Corpus annotation, sometimes referred to as 'tagging', can be thought of as a process of labeling a corpus with additional information such as linguistic, pragmatic or semantic tags in order to locate specific events or phenomena within the corpus and make it more useful for a given goal (Hovy and Lavid, 2010). Such annotation can be carried out by hand or automated through use of machine learning or other machine-based approaches. In the learning sciences, our study parallels Dönmez et al. (2005), who developed TagHelper technology for automating multi-dimensional analysis of collaborative learning data. This technology demonstrated its ability to accurately apply a predefined coding scheme, suggesting a notable reduction in manual effort and enhanced efficiency in data analysis for Computer-Supported Collaborative Learning (CSCL) research.

A blended inductive approach to corpus annotation

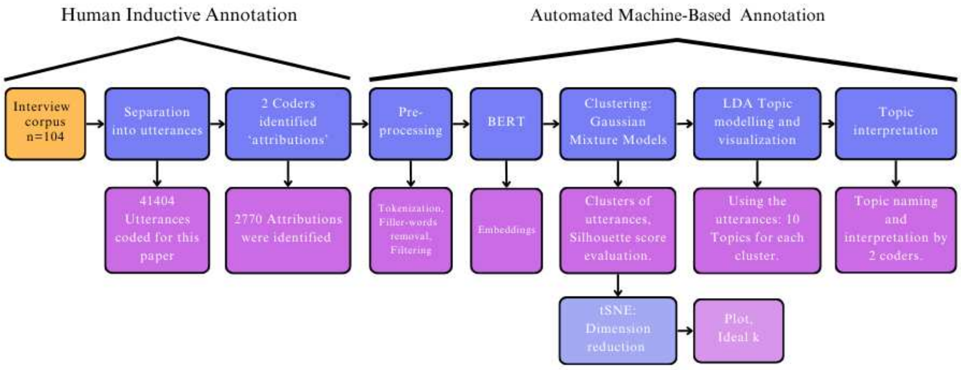
With the benefits of inductive annotation in mind, our team has developed a blended inductive annotation pipeline that blends inductive human coding with current approaches to machine learning. The immediate objectives in doing so are to cluster the various outcome statements made by interview participants, and associate each statement within a given cluster with a set of keywords or topics that are produced from one or more topic models. The resulting cluster labels and keywords then serve to annotate the corpus by indicating which outcome statements may be reliably grouped together and the potential topical terms that may be associated with them.

The inductive human annotation process was initially applied to the corpus of interviews. It led to a labeled data set in which individual sentences in participant interviews were labeled when they carried information about an outcome of the participants' arts participation. The coding scheme for identifying and labeling outcomes within the corpus of interviews is described in detail in Corrigan et al. (2023).

Here, we pick up on the annotation process using the data that resulted from use of the coding scheme. After isolating the rows of the corpus that present one or more outcome statements, we follow a standard pipeline (Figure 1) for preparing the outcome statements for analysis; and then apply a set of unsupervised machine learning techniques to cluster the outcome statements into related groups. We then apply Latent Dirichlet Allocation, or LDA, on the text in each of the clusters in order to identify key words and candidate labels that can be associated with the resulting topics. The outcome statements are subsequently indexed by their cluster number and topic labels, facilitating further research on the corpus.

Data and methods

Figure 1
Summary of the Full Blended Inductive Annotation Process



Interviews making up the study's corpus were collected using a semi-structured retrospective interview protocol designed to encourage descriptions of outcomes interviewees experienced as a result of their participation in the arts and arts programs. The pool of participants represents a convenience sample drawn from Australia, the United Kingdom, and the United States. A total of 102 semistructured retrospective interviews were conducted between September 2021 and August 2022, and subsequently transcribed and coded. A subset of 24,227 rows of the corpus that were coded by the same two coders was selected for analysis. Interrater reliability for the data was high with a percent agreement of 94.25% and an estimated Gwet's AC1 of 0.9345. A total of 2,770 rows (11.4%) contained one or more outcome statements and were subsequently included in the analysis.

Data treatment

The data cleaning process followed an established pipeline for topic modeling and natural language processing in general (Pustejovsky & Stubbs, 2012). This involved removal of punctuation, tokenization, removal of stop words and stemming filler terms and phrases such as 'Erm', 'Um', 'Uh', 'Uhm', and terms such as 'Like', 'So', 'You know', 'So, like', 'So, yeah', 'Yeah', and 'Yes' when in the first position of a given sentence. If an utterance was left empty after this removal, it was excluded from the dataset. After cleaning, the dataset was reduced to 2,736 attributions, which were then used for further computational analysis. This step ensured that the data was in a suitable form for the analysis, increasing the reliability of the results obtained.

Clustering

Using the *Transformers* library (Wolf et al., 2020) in Python, BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art natural language processing (NLP) model, was employed to convert each sentence into a numerical embedding. BERT leverages deep learning and attention mechanisms to understand the context of words in a sentence, capturing both the semantic and syntactic nuances (Devlin et al., 2018; He et al., 2020). This results in a high-dimensional vector for each utterance, encapsulating its meaning in a form amenable to computational analysis. Upon acquiring the embeddings, we applied the *Scikit-learn* library (Pedregosa et al., 2011) in Python, and employed its Gaussian Mixture Model (GMM) clustering algorithm, an advanced variant of k-means clustering. GMM not only determines the central point of a cluster but also models the distribution of the data, allowing for the identification of clusters with varying shapes, sizes, and densities. This adaptability makes GMM particularly suitable for complex datasets with intricate structures. We iterated the clustering process varying the number of expected clusters, $k = \{3 : 20\}$, in order to explore the potential number of groups present in the data.

To visualize and interpret the high-dimensional clusters generated by GMM, we employed t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten & Hinton, 2008), a dimensionality reduction technique recognized for preserving the local structure of the data. t-SNE reduces the dimensions of the data while maintaining the relative distances between points, resulting in a two-dimensional scatter plot that reveals the natural groupings and separations in the data. This visualization serves as a tool for intuitively understanding the underlying patterns and relationships within the textual data. The silhouette score was estimated at each value of k as an indication of how well the clusters could be separated as the number of clusters was varied.

Topic modeling

In the second phase of the analysis, the Gensim library (Řehůřek & Sojka, 2010) in Python was used to identify topics within each of the clusters using Latent Dirichlet Allocation (LDA). A topic in LDA is a multinomial distribution over the terms in the vocabulary of a given corpus. To interpret a topic, it is typical to examine a ranked list of the most probable terms for that topic, typically using three to thirty of the most prevalent terms from the list.

As noted consistently in the literature on topic modeling, associating meaning to topics is difficult within the LDA framework (Chang et al., 2009, e.g.). In order to support the effort of interpreting the resulting topics, the LDAvis library (Sievert & Shirley, 2014) in Python (Mabey, 2021) was utilized to generate a series of interactive visualizations and metrics that the research team used to a) agree on the meaning of each topic, b) investigate the prevalence of the various topics, and c) better understand how the topics relate to one another.

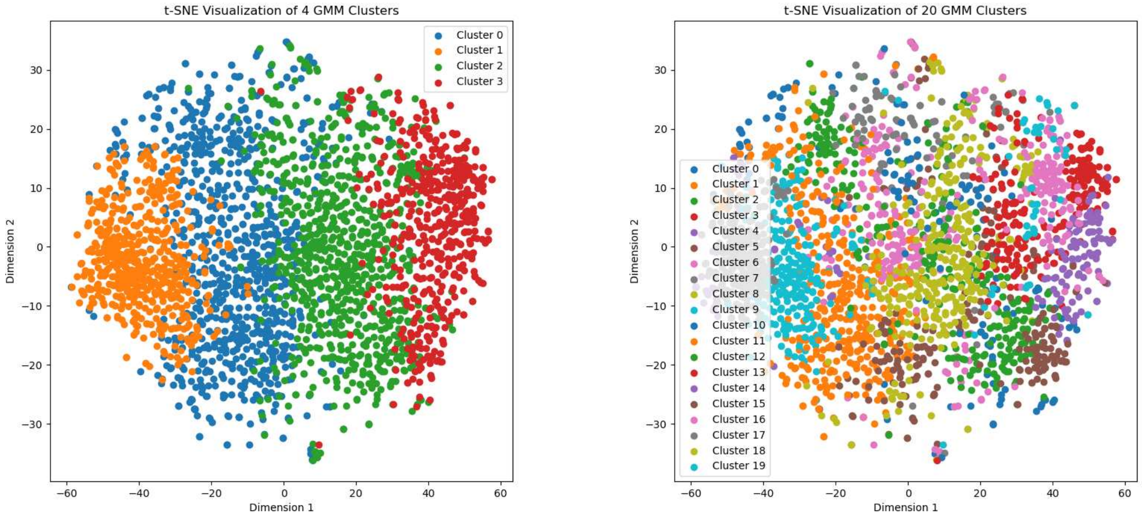
Results

We effectively applied Gaussian Mixture Model (GMM) clustering to our dataset. Our iterative process explored a range of cluster counts, specifically k-values from 3 to 20, to determine the optimal number of clusters. During each iteration, we calculated a silhouette score, which measures the cohesion and separation of data within clusters. A high silhouette score indicates well-defined clusters, while a lower score suggests clusters that are overlapping or less distinct. We observed an elbow point at k=4, where the silhouette score reached its peak at 0.212. This score, albeit modest, suggested a reasonable balance between cluster cohesion and separation at this point. However, in line with existing literature (Weston, Shryok & Fisher, 2023, e.g.), we recognized the importance of human interpretability in the clustering process. This decision underscores the significance of integrating human judgment in the analysis, especially when dealing with complex and nuanced datasets such as interview transcripts. The clusters, while less cohesive from a measurement standpoint, yielded richer insights when interpreted by members of the research team.

Figure 2 presents two of the t-SNE plots resulting from this stage of the pipeline. Each colored point represents an individual outcome statement from the corpus. In the first instance, where k was set to four clusters, the boundaries between the four groups of outcome statements are relatively clear, though there are individual outcome statements that cross over into other clusters. When one examines the second cluster, where k was set to 20, clusters of outcome statements are less well differentiated as one would expect from the related silhouette score.

Upon applying Latent Dirichlet Allocation (LDA) to the clusters derived from our GMM analysis, we identified 10 distinct topics in each cluster, each marked by its own set of keywords. Table 1 provides a sample of the full set of annotations that result from our blended inductive approach. Each row of the table contains an outcome statement that was identified though inductive coding carried out by the research team. Each outcome statement is indexed by its cluster and within-cluster topic resulting from the unsupervised clustering process and topic modeling process, respectively. In the fourth column, the outcome statements are further annotated with the ten key words associated with each topic.

Figure 2
Sample t-SNE Plots of Clusters of Outcome Statements with k = 4, 20



t-SNE Plot, k = 4 t-SNE Plot, k = 20

To delve deeper into these findings, we involved two coders from our initial annotation team for further interpretation. They were tasked with labeling these topics based on the provided utterances and keywords. Samples of the coders' category labels are given in the fifth and sixth columns of Table 1. Their responses revealed a high degree of congruence, albeit with subtle differences in their perspectives. For example, one topic was named 'Transitioning to professional arts experiences' by one coder and 'Shifting participation in the arts' by another, reflecting similar interpretations. Other topics were similarly aligned, such as 'Support from teachers, mentors, others' and 'Sustained support'. A notable variation appeared in one instance, as seen in row three of Table 1, where the topic was labeled 'Mental, physical, spiritual health' by one coder and 'Noticing in your context' by another. Rather than viewing this as a discrepancy, we consider it an intriguing example of how different researchers can assign coherent, yet distinct, interpretations to the same data. This underscores the multifaceted nature of our dataset and the value of integrating human judgment with computational analysis in revealing the diverse experiences and outcomes of arts participation. Importantly, such annotation systems can be personalized to specific researchers or teams in order to improve the usability of the annotation system for specific individuals or groups.

Discussion

In response to the historical context and growing interest in understanding the diverse impacts of arts participation, our demonstration study suggests a methodology that is aligned with and potentially expands existing approaches. Through its emphasis on induction, the annotation process, which integrates human interpretation with machine learning, supports open exploration of the multifaceted impacts of arts participation.

Our experience with the blended inductive approach described here, suggests it may be fruitful for investigating outcomes of arts participation that go beyond well recognized types or dimensions. This tentative claim is supported by the alignment of our identified topics with established themes in arts research. For instance, the congruence in coder interpretations, ranging from professional development in the arts to mental health implications, parallels themes prevalent in existing studies (e.g., Guhn et al., 2020). However, it is important to approach these findings with a degree of caution, recognizing that as an exploratory study, our results primarily serve to open the conversation and lay the groundwork for more in-depth research. This approach offers a potential pathway for further exploration and validation in the broader field of the learning sciences, encouraging a continued and evolving dialogue on the impacts and outcomes of arts participation.

Table 1
Sample t-SNE Plots of Clusters of Outcome Statements with k = 20

Utterance	Cluster (k)	Topic	Keywords within the topic	Topic name by Coder 1	Topic name by Coder 2

"And I also like started facilitating all these workshops and stuff, and doing all these, ah, that is also when I started getting major commissions from councils and things like that as well. "	15	7	Started, older, start, gone, etiquette, schools, noticed, getting, professional, auditions.	Transitioning to professional arts experiences	Shifting participation in the arts
"Like there was there was the crazy support system you had mentors the program director was always around to ask, Like help you out to ask like answer questions and give you advice"	3	8	Actually, advice, came, crazy ,director, mentors, reason, program, help.	Support from teachers, mentors, others	Sustained support.

We consider that the potential of the approach for identifying diverse sets of outcome statements, along with information regarding their location within the corpus, could yield a resource for others working to better understand the full universe of outcomes of participation in the arts. This approach not only has the potential to enhance analyses but also helps to ensure that findings are grounded in the actual experiences and perspectives of participants. In terms of generalizability, our study serves as an instance of using ML and human annotation to support qualitative research in a corpus for the learning sciences. This approach, inspired by the integration of Grounded Theory Methodology (GTM) and Machine Learning (ML) as suggested by Muller et al. (2016), presents a model that can be adapted to other research areas within the learning sciences. Drawing connections to earlier work using automated corpus analysis in learning sciences such as Dönmez et al. (2005), our study uses unsupervised machine learning for efficient corpus annotation, balancing streamlined analysis with depth and accuracy. Our bottom-up, inductive method, supports analyses that are both data-driven and rich in contextual, participant-specific insights. It demonstrates how computational methods can complement traditional qualitative analysis, providing a more nuanced and comprehensive understanding of complex datasets.

Looking forward, our next step involves employing a supervised learning approach to train a classifier that can predict whether a given utterance in this corpus of interviews is an attribution in the arts or not. This progression mirrors the structure proposed by Nelson (2020) in her framework of computational grounded theory, adapted to the context of the learning sciences. Nelson advocates for the incorporation of computational methods into inductive analyses within the social sciences, arguing for the enhanced efficiency, reliability, and validity these methods bring. Similarly, our future efforts aim to leverage these computational techniques in the learning sciences to enrich content analysis, ensuring that our methods remain grounded in the disciplinary knowledge of the field. This approach not only supports further qualitative analyses, but also stands to provide insights into the prevalence and representativeness of patterns within the larger corpus.

References

Belfiore, E., & Bennett, O. (2007). Rethinking the social impacts of the arts. *International journal of cultural policy*, *13*(2), 135-151. https://doi.org/10.1080/10286630701342741

Betts, J. D. (2006). Multimedia arts learning in an activity system: New literacies for at risk children. *International Journal of Education & the Arts*, *7*(7).

Bingham, A. J., & Witkowsky, P. (2021). Deductive and inductive approaches to qualitative data analysis. *Analyzing and interpreting qualitative data: After the interview*, 133-146. https://doi.org/10.3102/1682697

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage.

Catterall, J. S. (2009). *Doing well and doing good by doing art*. Imagination Group/I-Group Books.

Catterall, J. S., Dumais, S. A., & Hampden-Thompson, G. (2012). *The arts and achievement in at-risk youth: Findings from four longitudinal studies*. National Endowment for the Arts.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3-21. https://doi.org/10.1007/bf00988593

Dadswell, A., Bungay, H., Wilson, C., & Munn-Giddings, C. (2020). The impact of participatory arts in promoting social relationships for older people within care homes. *Perspectives in public health*, *140*(5), 286-293. https://doi.org/10.1177/1757913920921204

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805. https://doi.org/10.18653/v1/N19-1423

Donmez, P., Rose, C. P., Stegmann, K., Weinberger, A., & Fischer, F. (2005). Supporting CSCL with Automatic Corpus Analysis Technology, to appear in the Proceedings of Computer Supported Collaborative Learning. Gries, S. T., & Berez, A. L. (2017). Linguistic annotation in/for corpus linguistics. *Handbook of linguistic annotation*, 379-409. https://doi.org/10.1007/978-94-024-0881-2_15

Guhn, M., Emerson, S. D., & Gouzouasis, P. (2020). A population-level analysis of associations between school music participation and academic achievement. *Journal of Educational Psychology*, *112*(2), 308–328. https://doi.org/10.1037/edu0000376

Halverson, E. R. (2013). Digital art making as a representational process. *Journal of the Learning Sciences*, *22*(1), 121-162. https://doi.org/10.1080/10508406.2011.639471

Halverson, E., & Sawyer, K. (2022). Learning in and through the arts. *Journal of the Learning Sciences*, *31*(1), 1-13.https://doi.org/10.1080/10508406.2022.2029127

Holochwost, S. J., Propper, C. B., Wolf, D. P., Willoughby, M. T., Fisher, K. R., Kolacz, J., Volpe, V. V., & Jaffee, S.R. (2017). Music education, academic achievement, and executive functions. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(2), 147–166. https://doi.org/10.1037/aca0000112

Hovy, E., & Lavid, J. (2010). Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, *22*(1), 13-36.

He, H., & Choi, J. (2020, May). Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with BERT. In *The Thirty-Third International Flairs Conference*. Ingwersen, 1992; Mabey, B. (2021). pyLDAvis documentation. Jindal-Snape, D., Davies, D., Scott, R., Robb, A., Murray, C., & Harkins, C. (2018). Impact of arts participation on children's achievement: A systematic literature review. *Thinking Skills and Creativity*, *29*, 59–70. https://doi.org/10.1016/j.tsc.2018.06.003

Kosma, M., Erickson, N., Savoie, C. J., & Gibson, M. (2020). The effectiveness of performative aerial practice on mental health and the love of movement. *Research in Dance Education*, 1–18. https://doi.org/10.1080/14647893.2020.1784868

Matarasso, F. (1997). Use or ornament. *The social impact of participation in the arts*, *4*(2), 34-41.

Merli, P. (2002). Evaluating the social impact of participation in arts activities. *International journal of cultural policy*, *8*(1), 107-118. https://doi.org/10.1080/10286630290032477

Muller, M., Guha, S., Baumer, E. P., Mimno, D., & Shami, N. S. (2016, November). Machine learning and grounded theory method: convergence, divergence, and combination. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work* (pp. 3-8). https://doi.org/10.1145/2957276.2957280

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3-42. https://doi.org/10.1177/0049124117729703

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88-102. https://doi.org/10.1177/0165551515617393

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pustejovsky, J., & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

Řehůřek, R., & Sojka, P. (2010, May). Software framework for topic modeling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45-50). Valletta, Malta: ELRA.

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70). https://doi.org/10.3115/v1/w14-3110

Simpson Steele, J. S. (2019). Where are they now? Graduates of an arts integration elementary school reflect on art, school, self and others. *International Journal of Education and the Arts*, *20*. https://doi.org/10.26209/ijea20n11

Sol, S. L., Gustrén, C., Nelhans, G., Eklund, J., Johannisson, J., & Blomgren, R. (2021). Mapping research on the social impact of the arts: what characterizes the field?. *Open Research Europe*, *1*. https://doi.org/10.12688/openreseurope.14147.1

Stevenson, L. M., & Deasy, R. (2005). *Third space: When learning matters*. Arts Education Partnership.

Stuckey, H. L., & Nobel, J. (2010). The connection between art, healing, and public health: A review of current literature. *American Journal of Public Health*, *100*(2), 254–263. PubMed. https://doi.org/10.2105/AJPH.2008.156497

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., & Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics*, *4*(1), 14-28. https://doi.org/10.2139/ssrn.3199324

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9*(11). https://doi.org/10.7717/peerj-cs.154/fig-3

Walker, E., Tabone, C., & Weltsek, G. (2011). When achievement data meet drama and arts integration. *Language Arts*, *88*(5), 365.

Wallis, S. A. (2007). Annotation, retrieval and experimentation. *Studies in Variation, Contacts and Change in English*, *1*.

Weston, S. J., Shryock, I., Light, R., & Fisher, P. A. (2023). Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *6*(2). doi:10.1177/25152459231160105

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (pp. 38-45). https://doi.org/10.18653/v1/2020.emnlp-demos.6

Acknowledgments

This research was funded by The Wallace Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Foundation. We would like to thank our community arts partners and participating alumni.