Information Research - Vol. 30 No. iConf (2025)

An ensemble framework for sentiment-embedded event evolution in diaspora oral archives

Jing Zhou

DOI: https://doi.org/10.47989/ir30iConf47338

Abstract

Introduction. Diaspora oral archives should be displayed in the context of diversity and inclusiveness rather than being glossed over by the dominant, normative group so their voices need to be spread further.

Method. This paper proposes an ensemble framework for sentiment-embedded event evolution in diaspora oral archives. It contains a knowledge representation model, an event evolutionary graph, and an event extraction workflow to extract entities, events, and relationships.

Results. The South Asian Oral History Project is selected as the data source. The key events, entities, event types, sentiment and event relations are extracted with natural language processing techniques to construct sentiment-embedded event evolutionary graph. Based on this, the event evolution, spatio-temporal and spatio-sentiment patterns are analysed.

Conclusion. Such methods allow researchers and archivists to engage in research on machine-assisted oral archives to ensure reproducibility, reduce interpretative biases, and efficiently and swiftly amplify hidden voices of '*the other*'.

Introduction

Diasporas' voices and stories are ignored in the mainstream public consciousness. Oral archives may communicate hidden voices, revealing alternative narratives to those published in written histories (Brinkhurst, 2012). In terms of data types, oral archives are the audio/video recordings with field notes and transcriptions collections. These interviews can take various shapes including autobiographical narratives by the interviewee, such as lifestyle interviews, or interviews with semi-structured, open-ended questions following a research agenda, such as thematic interviews, or collective information sessions, where many persons participate in the conversation (Thomson, 1998). Hence, diasporas oral archives always need to be read in relation to broader, dislocated contexts. The contexts include an account of events, personal affection, etc. Nevertheless, on the one hand, similar oral archive practices concentrate more on the collection and preservation processes, neglecting semantic features and dynamic correlation in stories so events in these projects are barely connected by now. On the other hand, the participants-generated stories are unstructured textual data that are not predefined and indexed so it's hard to integrate or retrieve entities and events from them. This calls for new models and processes that enable intuitive access to event-related knowledge in diaspora oral archives. Hence, this study will focus on events and sentiments in diaspora oral archives, together with related personal narrative elements. These events and their relations, as well as entities, are vital components of diasporas' stories, constituting basic knowledge units of diaspora oral archives.

Diaspora oral archives are usually highly personal and emotive stories. The sentiment in archives creates a new legibility of the individual narratives (Roeschley and Kim, 2019; Jones, 2019). This study introduces the concept of "sentiment-embedded events" to describe events within these narratives that are closely tied to emotional responses, particularly those evoked by specific locations and contexts. Using a sentiment analysis framework that categorizes emotions into positive, neutral, and negative, this research adopts a coarse-grained approach to analyse sentiments associated with places and events. While this study does not aim to capture the finer nuances of affective experiences, it provides an overarching perspective on the emotional patterns linked to immigration activities and events.

An ensemble framework for sentiment-embedded event evolution is employed in this study, including a knowledge representation model, an event evolutionary graph (EEG) and automatic event extraction workflow by natural language processing (NLP). Precisely, this research exploits the sentiment-embedded event evolutionary graph (SEEGraph) to interpret and analyse these events and their relations, including the sentiment entailed in events. Besides, summarizing diaspora oral archives through manual annotation can be very tedious (or nearly impossible for long texts or large corpora). Many tasks have been successfully adapted to machine learning models in NLP that can help humans extract and summarize information from text automatically, as well as enhancing consistency and scalability in data analysis. Hence, events and their relations, and sentiments will be extracted through NLP in this study.

Motivated by these ideas, we hope to answer the following questions in our research:

– What knowledge representation model can capture sentiment-embedded events and their relationships in diaspora oral archives?

– How can machine learning models be integrated effectively to extract and enrich information in the proposed knowledge representation model?

– What are the meaningful patterns and insights of sentimental trajectories and event evolution in diaspora narratives?

Related works

Some researchers began to pay attention to knowledge representation of diaspora oral archives by text mining and NLP. These latest technologies create unique possibilities for the analysis of oral history interviews (Pessanha and Salah, 2021), such as identifying key topics within the histories, including events, and social or political issues (Rieping, 2022; Brown and Shackel ,2023).

Recently, sentiment in diaspora archives has drawn some researchers' attention. For instance, the First Days Project (Caswell and Mallick, 2014), Harvest Moon Oral History and the Flin Flon Heritage Project provide opportunities to acquire, describe, and preserve affective records that recount affective emotions (Grant, 2020). As for the automatic identification of complicated sentiment in diaspora archives, sentiment analysis is a relatively new approach for retrieving affective information in contexts of diaspora oral histories and interviews (Dominguès et al., 2019; Ozdemir and Bergler, 2015).

To further extract and store complete events and their relations, a conceptual model and SEEGraph are introduced in this study to standardize knowledge representation and enhance affective narration. The EEG originates from the concept of the event knowledge graph, which emphasizes the realization of goals such as events logic mining, and generalization. Event knowledge graphs have previously been employed in areas such as news articles (Rudnik et al., 2019), and contemporary and historical events (Gottschalk et al., 2023). The SEEGraph is proposed to provide a foundational descriptive framework for the event extraction results in NLP, while also incorporating sentiment within the events.

Research method

Knowledge representation model

The classes, properties, and hierarchical structures in the CIDOC CRM ontology are reused to construct the ontology of diaspora oral archives. CIDOC CRM is a high-level ontology that could play the role of the mediated schema in the information integration of cultural heritage data and their correlation with library and archive information (Bekiari et al., 2022). In CIDOC CRM classes like E53 Place, E2 Temporal Entity, E5 Event, E39 Actor, E55 Type are always connected and represented with an Event. The ontological framework of diaspora oral archives is shown in Figure 1. Entities in the ontology include Event, Event Type, Action, Participant, Time, Location, and Sentiment. Relations between Events are temporal relation and causal relation. As for relations between entities, this research reveals spatio-temporal relations and spatio-sentiment relations.

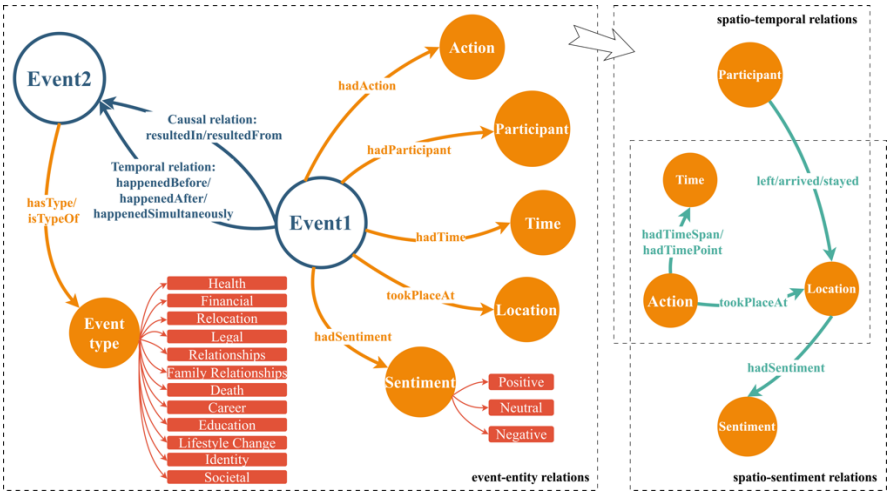

Figure 1. The ontological framework of diaspora oral archives

Event evolutionary graph

Since the EEG is event-centered, including two types of nodes (events and entities) and three types of directed edges (between events, between events and entities, and between entities and entities). The structure of the EEG is formally expressed as G, as follows:

$$G = \{(s, r, o) \mid \{s, o\} \in N, r \in P, N = N_{evt} \cup N_{ent}, P = P_{evt-evt} \cup P_{evt-ent} \cup P_{ent-ent}\}$$

Here the nodes in the EEG are represented as N, mainly including event nodes N_{evt} and entity nodes N_{ent}. Meanwhile, the edges in the EEG are represented as P, mainly including relations $P_{evt-evt}$ between events, relations $P_{evt-ent}$ between events and entities, and relations $P_{ent-ent}$ between entities. In the EEG, event nodes, entities nodes, and their relations are determined by the knowledge representation model.

(1) Events.

An event refers to an objective event or state change consisting of one or more action characteristics participated by one or more arguments in a specific time period or a specific region. Considering the ontological structure, the event is formally expressed as follows:

$$e = (A, P, T, L, S, ET)$$

Here, 'e', 'A', 'P', 'T', 'L', 'S', 'ET' represents Event, Action, Participant, Time, Location, Sentiment and Event Type.

(2) Event relations.

The causal and temporal relations are the most common relations between events (Caselli and Vossen, 2017). Temporal relations can be defined as before, after, simultaneous, begins, ends, etc. The causal relation can be described in precise semantics or hidden in logic, which means it could be inferred (Liu et al., 2021).

Event extraction workflow

We construct an event extraction workflow to construct the information as described in the knowledge representation model. There are four main components for the EE workflow:

(1) Key events and entities extraction.

The real-world events have different granularities, from the top-level themes to key events and then to event mentions corresponding to concrete actions (Zhang et al., 2022). In this context, we propose a new task, key event detection at the intermediate level, which aims to detect event blocks. Each archive is divided into several event blocks manually according to contents. Considering the spatial role that places play and the study's objectives, entities labeled as 'GPE' (Geo-Political Entity) are extracted from each event block. Events associated with placenames are identified as key events. Additionally, other entities related to specific locations, including Participant, Action, and Time, are also extracted.

(2) Event classification.

The event type classification task is based on a taxonomy and automatic classifier. The Major Life Events Taxonomy is employed as the taxonomy of diaspora key events in this research. It is a U.S.-based list of major life changes that people experience (Haimson et al., 2021). In this taxonomy, '*life events*' are used as an umbrella term to encompass life experiences involving both moments and processes of change. It can be used to classify diasporas' key events in their lives into 12 categories: Health, Financial, Relocation, Legal, Relationships, Family Relationships, Death, Career, Education, Lifestyle Change, Identity, and Societal.

(3) Sentiment analysis.

The connection between sentiment and space is particularly significant, as certain locations often evoke distinct emotional responses. Sentiment analysis concerning places can reveal their intuitive feelings about their immigration activities and events. Thus, Aspect-based sentiment analysis (ABSA) is applied to analyze the sentiments ('*positive*', '*negative*' and '*neutral*') of diasporas with the aspects of places (Syamala & Nalini, 2019).

(4) Event relation extraction.

The purpose of event relation extraction is to extract relative clauses and relationship indicator words in the corpus based on syntactic relationships and matching rules, which are tuples in the form of <event A, relation, event B>. This article chooses the commonly used pattern-matching method to extract event relations. It summarizes event relation words and sentences with explicit conjunction (Table 1) and uses a predefined rule base to perform semantic relationship matching.

Relations	Properties	Syntactic pattern	Conjunction
Casual relation	resultedIn/ resultedFrom	{Event_effect} <Conj> {Event_cause}	because, because of, since, as, for, etc.
		{Event_cause} <Conj> {Event_effect}	because of this, so, thus, therefore, consequently, etc.
		<Conj> {Event_effect} {Event_cause}	another important factor / reason of
		<Conj> {Event_cause} {Event_effect}	because, because of, since, as, for, etc.
Temporal relation	happenedBefore/ happenedAfter/ happenedSimulta-neously	{Event_previous} <Conj> {Event_latter}	next, then, after that, later, etc.
		{Event_latter} <Conj> {Event_previous}	previously, prior to this
		{Event_simul1} <Conj> {Event_simul2}	simultaneously
		<Conj>{Event_latter} {Event_previous}	before
		<Conj>{Event_previous} {Event_latter}	after

Table 1. The syntactic pattern and conjunction of event relation pattern

Analysis and results

Data source

The author analyses 42 English transcripts of The South Asian Oral History Project (SAOHP) at the University of Washington Libraries. The SAOHP represents one of the first attempts in the U.S. to record pan-South Asian immigrant experiences in the Pacific Northwest using the medium of oral history. The SAOHP is marked by key historical events that drew South Asians to the United States. These interviews include important events in diasporas' lives reflecting religious, linguistic, occupational, and gender diversity and provide rich insight into the changing experiences of South Asians in the Pacific Northwest. The interviews are in English digitalized and transformed into textual transcripts. Each interview lasts more than 60 minutes and the transcript contains over 20 pages of text. These transcripts were formatted as CSVs separated into chunks of text.

Model evaluation

The spaCy library, DeBERTa and GPT3.5 are selected to conduct the event extraction workflow: entity extraction, event type extraction, sentiment analysis, event relation extraction. Precisely, the spaCy library, and GPT3.5 are conducted on entity extraction and event relation extraction. While in terms of the limitation of the spaCy library, only GPT3.5 are applied to event type extraction and sentiment analysis. Besides, a DeBERTa-based ABSA classifier is also utilized as a supervised machine learning method in sentiment analysis. About 900 manually annotated

samples, are used as the baseline for the evaluation. The performance of these models is assessed using standard NLP metrics such as precision, recall and F1 score. The evaluation results are shown in Table 2.

Tasks	Methods	Evaluation metrics		
		Precision	Recall	F1
Manual annotation		1	1	1
Entity extraction	The spaCy library	42.37%	52.49%	46.89%
	GPT3.5	74.26%	86.80%	80.04%
Event type extraction	GPT3.5	69.75%	77.49%	73.42%
Sentiment analysis	GPT3.5	89.38%	92.39%	90.86%
	DeBERTa	93.54%	92.39%	92.96%
Event relation extraction	The spaCy library	30.25%	35.25%	32.56%
	GPT3.5	12.42%	10.49%	11.37%

Table 2. The evaluation of chosen models

The results demonstrate GPT-3.5 achieves the highest performance in the entity extraction task, while it also demonstrates good performance in event type extraction. For sentiment analysis, the DeBERTa-based ABSA model delivers the best results. However, for event relation extraction, manual validation and complementary methods remain necessary to ensure accuracy.

For each task, the model with the best performance will be utilized to complete all extraction and classification processes. And the extracted data is checked and supplemented through manual inspection. There are 1533 key events determined altogether. And a total of 1043 temporal relationships and 632 causal relationships are ultimately identified.

SEEGraph construction

The proposed event extraction workflow is applied to extracted defined entities, events, and relations, which are imported into the Neo4j database to realize the mapping of knowledge framework and ontology classes to properties and instances. To sum up, there are 1533 event nodes N_{evt}, 19435 entity nodes N_{ent}, 1675 relations $P_{evt-evt}$ between events, 25920 relations $P_{evt-ent}$ between events and entities, and 14825 relations $P_{ent-ent}$ between entities. Hence, the SEEGraph is generated which provides a structured framework for representing and reasoning about diasporas' events.

Event evolution analysis

The SEEGraph traces the sequences, patterns, and trajectories of events influencing diaspora movements, transformations, and experiences over time. For example, Seattle had its large share of Indians who were predominately engineers who worked for Boeing in the 1960s and 1970s. The job offers from Boeing have become a key facilitator for diasporas to immigrate during that period. As for AM (Figure 2), after getting a job at Boeing, he moved to Seattle. He made friends with colleagues who are also Indian diasporas and was admitted to the engineering college which collaborated with Boeing company.

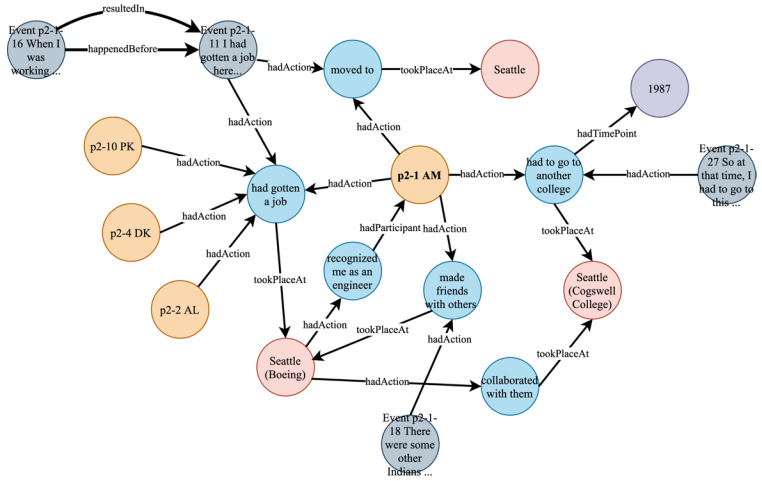

Figure 2. The event evolutionary process according to job offers from Boeing

Spatio-temporal analysis

The spatio-temporal analysis creates a multidimensional representation of spatial movements and temporal changes. The whole picture of interviewees' immigration activities is revealed in Figure 3. The nodes are the places they have stayed for more than one year and the connecting curves show the flows and years of immigration. We can conclude four types of immigration modes: single-directional, round trip, multi-directional, and internal trip.

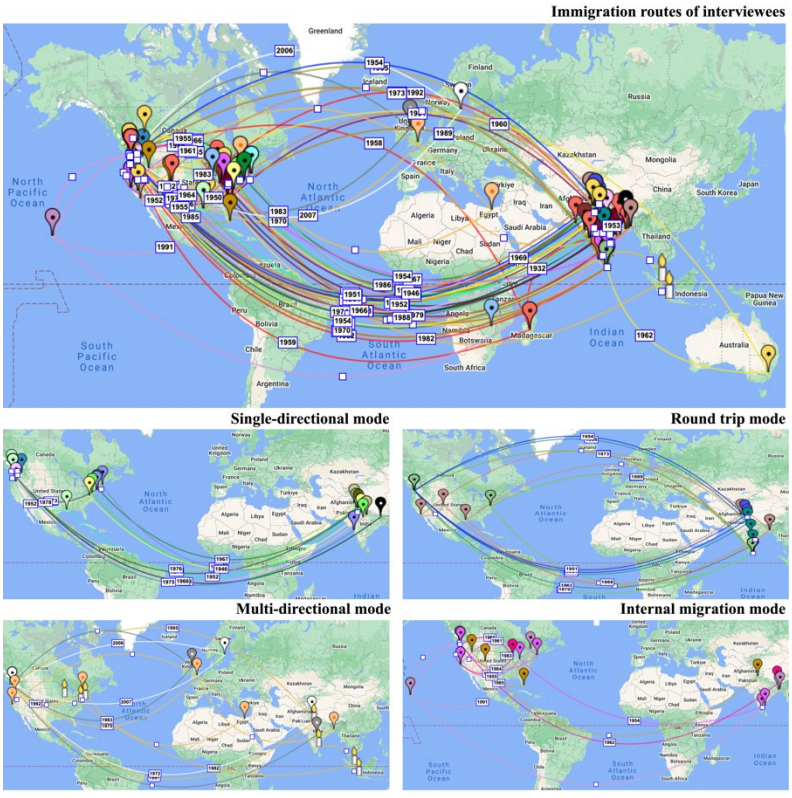

Figure 3. Interviewees' immigration routes and modes

Spatio-sentiment analysis

The spatio-sentiment analysis creates a rich and context-aware representation of the diasporas' sentiment towards geographical entities, especially their destination and homeland. As is shown

in Figure 4, In terms of their sentiment toward the most frequently mentioned locations, their comments on the USA and American cities are one-sided, and positive perspectives are much more than negative ones. According to their sentiment towards their homeland, the feelings are more complicated and conflicted. An almost equal quantity of positive and negative opinions indicates a mixed reviews attitude.

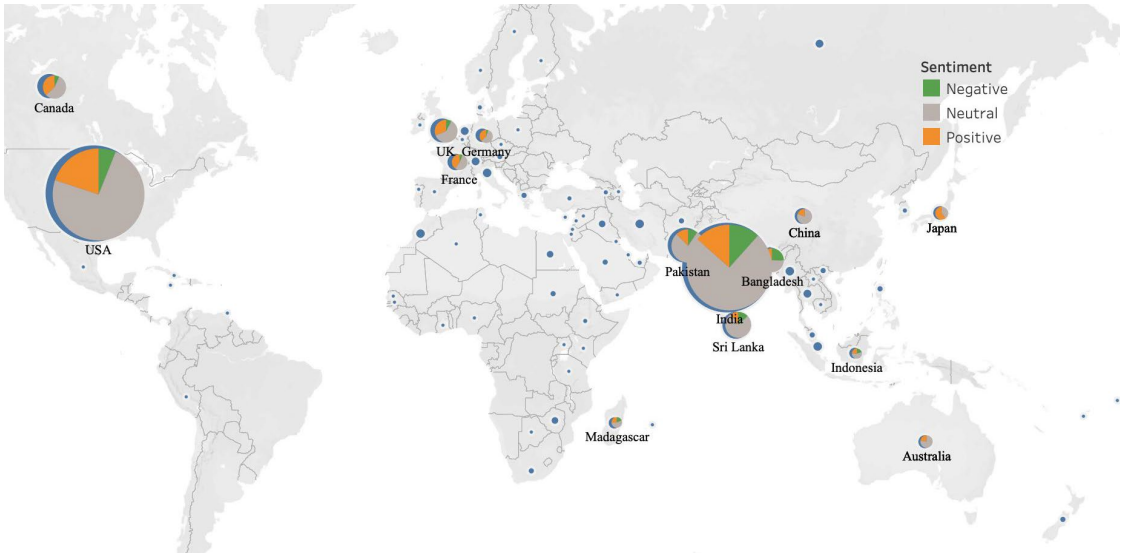

Figure 4. The diasporas' sentiment toward main locations

Discussion and conclusion

The analysis of diaspora oral archives serves as a pathway to represent diversified voices in repositories. In this research, the ensemble framework is introduced not only to realize the semantic, correlation, and structured expression of knowledge units such as events, entities, and attributes but also to clearly reveal the dynamic evolution rules and patterns between historical and personal events. this study has innovations and contributions in the following aspects:

Firstly, unlike previous event extraction conducted using relatively high-quality data such as media articles (Li et al., 2018; Rudnik et al., 2019), the presented study processes the information extraction of diaspora oral archives transformed into structured text. Specifically, the dataset used in this study is differentiated in the following aspects: (1) longer and complicated context in the interview conversation with imbalanced length, (2) complex event content and relations due to interviewees' different narratives, and (3) including many idiomatic expressions, slang, colloquialisms, or repeated expressions in oral narratives. To solve this, the ontological model is proposed to normalize the domain of diasporas' major life events. Instead of defining event as event trigger words and event arguments as in previous research (Guan et al., 2022), this study refines it as action, participant, time, location, sentiment, and event type. So, the extracted information in the event is more abundant and comprehensive. It fills research gaps in event detection which mainly focuses on political events or news corpora. Besides, it's the first time that the EEG combined with NLP pipelines to analyze diaspora oral archives. The performance of spaCy, LLMs, etc. has been examined in each task. The ensemble framework with multiple models is expected to be used in the processing of other diaspora oral archives and low-quality textual transcripts of oral materials in the future.

Secondly, the sentiment is involved, which is also first introduced into the EEG and diaspora oral archives. This study helps to reveal the reasons and results of the diasporas, how they are influenced and exert an influence on the mobility of the migration, and how they build up a sense of belonging toward the destination country. This led to intense confrontations, with both positive

and negative emotions. Though in this research the sentiment is coarse-grained, it can be seen as the first attempt to understand and respond appropriately to diasporas' emotional reactions. With embedded sentiment, diasporas' affective assessment of places and events deepens the understanding of their immigration activities.

Acknowledgements

I would like to express my gratitude to Prof. Charles Jeurgens for his careful reading of the manuscript and his valuable feedback and suggestions.

The oral history transcripts used in this study are part of the South Asian Oral History Project (SAOHP) collection at the University of Washington Library. While these materials are publicly accessible, their copyright remains with the creators (interviewees and interviewers). This study uses the data strictly for academic research purposes in compliance with the principles of '*Fair Use*.' The data has been anonymized, and no identifiable personal information is disclosed. Any results or findings are derived from analytical methods and do not reproduce or distribute the original content of the transcripts.

About the author

Jing Zhou is doctoral candidate in the School of Information Management, Wuhan University. Her research interest focuses on digital humanities, knowledge organization and natural language processing. She can be contacted at zhoujingwinky@whu.edu.cn.

References

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023. https://doi.org/10.48550/arXiv.2302.04023.

Bekiari, C., Bruseker, G., Doerr, M., Ore, C. E., Stead, S., & Velios, A. (2021, April). Volume A: Definition of the CIDOC conceptual reference model. Version 7.1. (https://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_7.1.2.pdf).

Brinkhurst, E. (2012). Archives and Access: Reaching Out to the Somali Community of London's King's Cross. Ethnomusicology Forum, 21(2), 243–258. https://doi.org/10.1080/17411912.2012.689470

Brown, M. & Shackel, P. (2023). Text Mining Oral Histories in Historical Archaeology", International Journal of Historical Archaeology, 27, 865–881. https://doi.org/ 10.1007/s10761-022-00680-5.

Caswell, M. & Mallick, S. (2014). Collecting the easily missed stories: digital participatory microhistory and the South Asian American Digital Archive. Archives and Manuscripts, Taylor & Francis, 42(1), 73–86. https://doi.org/10.1080/01576895.2014.880931.

Caselli, T. & Vossen, P. (2017, August). The event storyline corpus: A new benchmark for causal and temporal relation extraction. In Proceedings of the Events and Stories in the News Workshop (pp. 77-86). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2711.

Dominguès, C., Jolivet, L., Brando, C. & Cargill, M. (2019). Place and Sentiment-based Life story Analysis. From the Spanish Republican Army to the French Resistance. Revue Française Des Sciences de l'information et de La Communication, 17. https://doi.org/10.4000/rfsic.7228.

Gottschalk, S., Kacupaj, E., Abdollahi, S., Alves, D., Amaral, G., Koutsiana, E., ... & Thakkar, G. (2023). OEKG: The open event knowledge graph. arXiv preprint arXiv:2302.14688. https://doi.org/10.48550/arXiv.2302.14688.

Grant, K.A. (2020). Affective Collections: Exploring Care Practices in Digital Community Heritage Projects (Unpublished master thesis). University of Alberta, Alberta, Canada

Guan, S., Cheng, X., Bai, L., Zhang, F., Li, Z., Zeng, Y., ... & Guo, J. (2022). What is event knowledge graph: A survey. IEEE Transactions on Knowledge and Data Engineering, 35(7), 7569-7589. https://doi.org/10.1109/TKDE.2022.3180362.

Haimson, O.L., Carter, A.J., Corvite, S., Wheeler, B., Wang, L., Liu, T. & Lige, A. (2021). The major life events taxonomy: Social readjustment, social media information sharing, and online network separation during times of life transition. Journal of the Association for Information Science and Technology, 72(7), 933–947. https://doi.org/10.1002/asi.24455.

Jones, M. (2019). Archiving the trauma diaspora: Affective artifacts in the higher education arts classroom. Marilyn Zurmuehlen Work. Pap. Art Educ, 2019, 1-14.

Li, Z., Ding, X., & Liu, T. (2018, July). Constructing narrative event evolutionary graph for script event prediction. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, Sweden (pp. 4201-4207). AAAI Press.

Liu, Y., Tian, J., Zhang, L., Feng, Y., & Fang, H. (2021). A Survey on Event Relation Identification. In Knowledge Graph and Semantic Computing: Knowledge Graph and Cognitive Intelligence: 5th China Conference, CCKS 2020, Nanchang, China, November 12–15, 2020, Revised Selected Papers (pp. 173-184). Springer Singapore. https://doi.org/10.1007/978-981-16-1964-9_14.

Ozdemir, C. & Bergler, S. (2015). CLaC-SentiPipe: SemEval2015 subtasks 10 B, E, and task 11. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (https://aclanthology.org/S15-2081.pdf).

Pessanha, F., & Salah, A. A. (2021). A computational look at oral history archives. ACM Journal on Computing and Cultural Heritage (JOCCH), 15(1), 1-16. https://doi.org/10.1145/3477605.

Rieping, H.A. (2022). Audio Segmenting and Natural Language Processing in Oral History Archiving. (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Massachusetts, U.S.A.

Roeschley, A. & Kim, J. (2019). “Something that feels like a community”: the role of personal stories in building community-based participatory archives. Archival Science, 19, 27–49. https://doi.org/10.1007/s10502-019-09302-2.

Rudnik, C., Ehrhart, T., Ferret, O., Teyssou, D., Troncy, R., & Tannier, X. (2019, May). Searching news articles using an event knowledge graph leveraged by wikidata. In Companion proceedings of the 2019 world wide web conference (pp. 1232-1239). Association for Computing Machinery.

Syamala, M., & Nalini, N. J. (2019). A deep analysis on aspect-based sentiment text classification approaches. International Journal of Advanced Trends in Computer Science and Engineering, 8(5), 1795-1801.

Thomson, A. (1998). Fifty years on: An international perspective on oral history. Journal of American History, 85(2), 581. https://doi.org/10.2307/2567753.

Yang, H., Zeng, B., Xu, M. & Wang, T. (2021). Back to Reality: Leveraging Pattern-driven Modeling to Enable Affordable Sentiment Dependency Learning. ArXiv Preprint. (https://www.researchgate.net/profile/Heng-Yang-

17/publication/355391949_Back_to_Reality_Leveraging_Pattern-driven_Modeling_to_Enable_Affordable_Sentiment_Dependency_Learning/links/6189682107be5f31b7590ae3/Back-to-Reality-Leveraging-Pattern-driven-Modeling-to-Enable-Affordable-Sentiment-Dependency-Learning.pdf).

Zhang, Y., Guo, F., Shen, J., & Han, J. (2022, August). Unsupervised key event detection from massive text corpora. In Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining (pp. 2535-2544). Association for Computing Machinery. https://doi.org/10.1145/3534678.3539395.