Indiana University | Kahyun Choi, Ph.D.

Unbiased AI for Poetry Analysis: Toward Equitable and Diverse Digital Libraries

1. Project Justification

The ever-growing large-scale digitized collections in libraries, archives, and museums (LAM) has greatly improved accessibility of remote resources for patrons. Meanwhile, the overwhelming amount of data necessitates more careful curation, intuitive browsing interfaces, and comprehensive annotations for patrons to use, which are difficult tasks for humans to achieve manually. These emerging challenges require large-scale data comprehension, such as artificial intelligence (AI) with a human-level understanding of literature. Consequently, embracing rapidly evolving AI technology is an urgent task for LAMs (Cordell, 2020). This project aims to improve the AI models' understanding of poetry, one of the most challenging types of text for AI to understand. In particular, the project team will tackle the algorithmic biases inherent in AI models that are naïvely trained from imbalanced datasets. Instead, we propose a new AI model development pipeline, with which AI models can accurately extract high-level metadata from poems, such as theme and emotion, including from obscure or underrepresented poems. We will achieve this goal via a close collaboration among digital librarians, the HathiTrust Research Center (HTRC), poetry experts, and AI-based natural language processing (NLP) researchers. As a result, the project team will develop an open-source toolkit to help librarians freely curate diverse and unbiased poetry collections from their libraries. In addition, we will also develop a poetry discovery system based on emotion and theme metadata to improve the public's accessibility to underrepresented poems. The project aims at fulfilling Program Goal 2 - Objective 2.3, which is supporting long-term research and career development of untenured tenure-track library and information science faculty.

1.1. AI to increase the accessibility of large-scale digital poetry collections

AI-based metadata extraction systems can increase the accessibility of massive digitized data collections as advanced AI models are better able to accurately capture descriptive metadata (beyond author names and titles) to an unprecedented level (Han et al., 2003; Bainbridge et al., 2014). As a result, patrons could browse or search from extensive data with high-level content-driven queries, such as "upbeat and romantic poetry written in the 19th century," where emotion and theme-related keywords add an additional interface. If the AI-driven metadata becomes as accurate as manual labels in the future, librarians would be able to curate collections and give patrons the freedom to discover items beyond the current full-text search services. In addition, the large-scale metadata allows digital humanities scholars to obtain insight from large-scale data, while previous research tends to focus on smaller sets of data (Fenlon, 2014; Kaplan, 2015; Underwood, 2019). The advantage of such automatically derived metadata is obvious: as manual annotation is often too expensive, and the size of ever-growing digital collections is gigantic, only a relatively small number of collections with limited types of manual metadata are currently available. As of now, searching and browsing systems are either utilizing only part of the collections or providing simple metadata, e.g., author names and titles. Hence, hard-to-find items stay invisible and are marginalized due to a lack of meaningful metadata (Heidorn, 2008). In this circumstance, AI-based metadata extraction systems can annotate the forgotten data with high-level descriptive information.

Despite their advantages, large AI models have also caused emerging concerns about their biasedness. For example, in natural language processing (NLP) applications, automatic annotation algorithms are often based on huge artificial neural networks, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) or Generative Pre-trained Transformer 3 (GPT-3) (Brown et al., 2020). While they reach unprecedented levels of comprehension with certain kinds of text, such as news articles and product reviews, more research is required to build AI models that understand literature. Poetry is considered the most difficult-to-understand type of literature even for humans due to the challenge brought by figurative language and multiple layers of meanings. Thus, creating machines

that understand poetry is a daunting task; attempting to have them catch up with high-level human intelligence will push the limit of AI, resulting in a great ripple effect within the context of NLP and literature analysis.

In this project, we focus on poetry as a representative of a marginalized literature type that can benefit from fair annotation by AI methods. After undergoing a downhill trend in popularity (Perelman, 1996), poetry has lately regained significant attention. For example, after Amanda Gorman's reading of her poem at the 2020 U.S. presidential inauguration, "traffic to Poets.org spiked dramatically, with 250% more individuals coming to the site than on the same day last year," according to the poetry collection website Poets.org. Furthermore, traffic on poetry websites increased significantly during the COVID-19 pandemic because it positively affected mental health during such a difficult time (Acim, 2021). It has long been known that literature is helpful for one's well-being in general (Hynes, 2019). "The place for a cure for the soul" has been inscribed in the great library of Alexandria in ancient Egypt, and Aristotle in *Poetics* emphasized the healing effect of literature through catharsis. Bibliotherapy, as an active use of literature for therapeutic purposes, has been performed for decades, with poetry as one of the main genres (Harrower, 1972; Mazza, 2016). We envision that diverse metadata can help organize massive digital poetry collections and become essential in this golden time of poetry, increasing accessibility to various poems already in collections.

High-level metadata for poetry and other difficult literature types, such as those addressing emotions and complex themes, is currently insufficient. Although poetry often uses emotional language, emotion is rarely annotated with poems, nor used for searching and browsing services in poetry websites and digital libraries (Khan et al., 2021). Meanwhile, AI's emotion recognition performance is not advanced enough for poetry because such sentiment analysis models have often excluded poetry from their training dataset due to the inherent difficulty in manually annotating poems (Haider, 2020). Thus, only a small number of emotion-annotated poems are available (Haider, 2020; Khan et al. 2021). Researchers only recently began to research how to define emotion taxonomy for poems (Kao, 2012; Jacobs, 2019; Saini, 2019; Haider, 2020). Compared to emotions found in poetry, themes found in poetry have been more widely explored, resulting in relatively larger-sized datasets (Lou, 2015; Navarro-Colorado, 2018). However, theme-labeled datasets are still insufficient when conducting large-scale deep learning research. Furthermore, since the definitions used to identify themes and information of the participating annotators are often unknown, it is necessary to systematize and build annotated poetry collections in a diverse, inclusive, and unbiased way.

1.2. Identifying and removing bias in AI models

We have witnessed how AI models are reinforcing social biases and stereotypes. For example, Google returned sexual items as search results of "black girls" but not "white girls" (Noble, 2018). In addition, facial recognition systems are known to be more accurate for men than women and for light-skinned than dark-skinned (Buolamwini & Gebru, 2018). Also, it was reported that GPT-3—the most representative foundation model on which NLP applications are based—showed bias against Muslims and other marginalized groups (Bommasani et al., 2021).

Data-driven AI models, such as deep learning models, are inherently prone to biases because the data collection process can be prejudiced or inherently biased. AI models often learn patterns and knowledge from a small amount of data curated by a small number of people, often from dominant groups. Because they do not represent the general population, the models tend to discriminate against marginalized people (D'Ignazio & Klein, 2020). There is a gender and racial imbalance in demographics in technology industries as well as LAMs, which can cause adversarial effects of AI toward marginalized groups. Also, large-scale data from the Internet often contain biases and stereotypes. These days, document discovery AI models are based on foundation models, such as BERT and GPT-3. They are first trained on bulk data, from Wikipedia, digitized books, to all documents on the Internet. Then, they are fine-tuned for the particular target problem. Because the initial raw data contain stereotypes and social biases, any models stemming from the foundation models amplify these biases unless additional action is done to avoid it (Bommasani et al., 2021; Bender et al., 2021).

Researchers, institutions, and governments have begun to understand the nature of bias, and have started creating guidelines to mitigate biases. The common ground inside and outside of LAMs at this moment is that biases can be mitigated only if we work actively toward equity at every stage of AI lifecycles. First, it is essential to put extra effort and attention into data collection and annotation (Jo & Gebru, 2020)—therefore, it is important to identify who is in charge of curating and annotating the data to determine if they represent user groups appropriately (D'Ignazio & Klein, 2020). Building collections and annotating data are critical "to benefit evenly or more to the historically most marginalized" (Bender et al., 2021). Second, during the iterative AI model-building process, tight collaborations with stakeholders, such as librarians and domain experts, are strongly recommended (Cordell, 2020). Third, the final AI model evaluation process needs to consider various user groups' responses in addition to overall model accuracies (Hu et al., 2017; Forde et al., 2021). Oftentimes, biased AI models yield higher accuracy by simply focusing on the performance of the majority. In this case, user evaluation can weed out biased AI models from the final model. Finally, to promote reproducibility and transparency, it is of pivotal importance to document every stage of the AI lifecycle, including subjective decisions, and then publish the documents along with open-sourced codes and data (Cordell, 2020; Schwartz et al., 2021).

1.3. PD's past research and long-term research agenda

The PD has experience with every stage of the workflow of AI models: from data collection and annotation, building AI models, to evaluation of the models. Her prior work identified and remedied critical issues in Music Information Retrieval (MIR), such as researchers' bias towards western music, a lack of user-centered evaluations and diversity in data, and gender imbalance in the research community. This research revealed how K-Pop, a non-western music genre, contributed to diversifying MIR research through 1) building genre taxonomy of K-Pop, 2) recruiting annotators of K-Pop collections from two groups, Koreans and Americans, and 3) analyzing how people from the two groups perceive K-Pop similarly or differently (Lee et al., 2013; Hu et al. 2014). She also participated in an informetric study on publications of the main MIR conference to increase awareness of gender imbalance in MIR (Hu et al., 2016). She helped build and host an event for holistic user-experience evaluations of AI models for music applications with the goal of convincing MIR researchers to focus more on users of the AI models (Lee et al., 2015; Hu et al., 2015). As for the AI models for music applications, she has developed automatic music annotation systems (Choi et al., 2014-2021), which analyzed song lyrics based on emotion and theme metadata using state-of-the-art AI technologies.

The PD's long-term research agenda centers around improving the accessibility of massive digital content from text to multimedia in LAMs. She will actively develop AI models that can extract high-level metadata from content and user-generated data. In this project, she shifts her focus from music and song lyrics to poetry collections. She also envisions that the transition can further lead to encompassing other bodies of literature and multimedia data in the long term. Most of all, the PD focuses on how to develop ethical AI models and build inclusive datasets that benefit all, including marginalized groups. Creating guidelines for other LAM researchers and librarians and publishing them for transparent communication make up a critical portion of her agenda. Moreover, the PD believes in the importance of adopting AI models in LAMs for use by all demographics within the general population. To that end, she is currently developing and offering library-based AI education programs to underserved youth to increase equity and accessibility (IMLS grant number: LG-250059-OLS-21).

2. Project Work Plan

2.1. Overview of Personnel and Collaboration with HTRC

The project team consists of PD Choi and a Ph.D. student. The PD will lead the intellectual and technical effort, oversee the entire project progress, and be responsible for the dissemination tasks. A Ph.D. student will assist

the PD in research and dissemination activities. The project team will consult with an Advisory Board (AB) which comprises five members with expertise in poetry, library and information science, digital scholarship librarianship, metadata generation, AI in libraries, large-scale digital libraries, and diversity and inclusion.

The project team will tightly collaborate with the HathiTrust Research Center (HTRC), a collaborative research center launched jointly by Indiana University (IU) and the University of Illinois at Urbana-Champaign (UIUC), along with HathiTrust Digital Library. HTRC enables the computational analysis of 17+ million digital items, a collection from more than 100 member libraries, including IU and UIUC libraries. They provide a mechanism to allow librarians and digital scholars to access copyright-protected data while complying with the US copyright law. We include HTRC as a digital library for our study, through which any librarian can access collections and use our AI models to utilize metadata-based curation. The PD has already established a collaborative relationship with HTRC—John Walsh, the director of HTRC, will guide collaboration with HTRC as an AB member.

2.2. Research Questions

The need to improve accessibility to massive digital collections necessitates a new, unbiased automated metadata extraction method. Hence, our goal is to achieve AI-based poetry analysis methods that robustly understand the semantics of poetry across a wide, all-inclusive range of sources. Throughout the project period, we aim at a sustainable and flexible ecosystem of AI models for rich metadata-based digital libraries, which is a common goal accomplished by librarians' and patrons' everyday activities. Our research questions are designed to address the aforementioned goals:

1) What are the underrepresented social groups within a digital collection of poems?
2) How can a potentially underrepresented poetry collection be accurately annotated based on the theme and emotion they convey?
3) How does an AI model analyze the poetry text and classify it into the different theme and emotion categories?
4) What is the best way to utilize auxiliary information about the poet and poetry in the AI-based automatic metadata extraction systems?
5) What kind of biases does an AI-based metadata extraction system for poems suffer from, especially those of underrepresented social groups?
6) How can we improve the performance of the AI-based poetry analysis systems regarding potential bias?
7) How can a librarian contribute to the pipeline of training, evaluating, improving, and using an AI model for poetry analysis?
8) How would such a system be easily used by many librarians at scale and continue to be impactful?

2.3. Overview of the Workflow

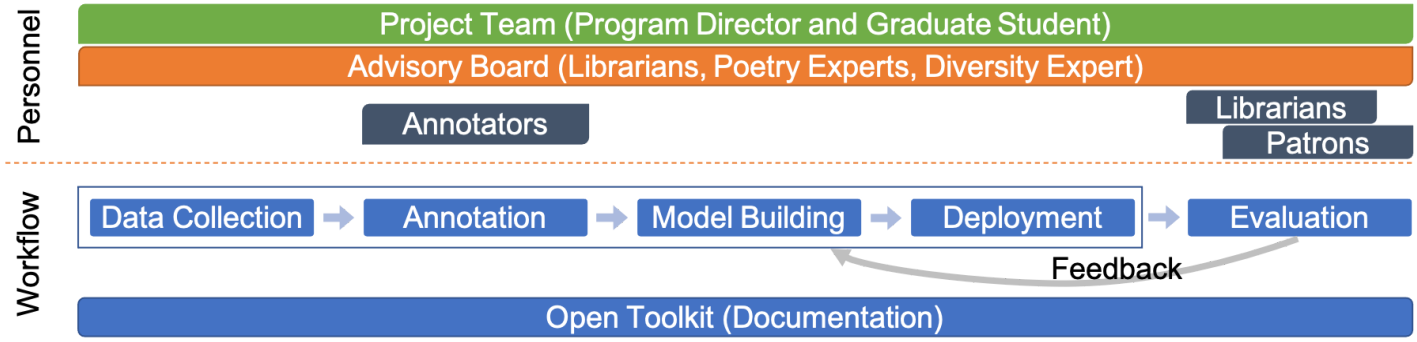

Fig. 1. The workflow for the proposed AI-based metadata extraction system. We involve human experts in every step of the workflow to reduce bias.

We designed the work plan to incorporate all the participating personnel's expertise into the AI-based poetry analysis system pipeline, as shown in Fig. 1. By doing so, the proposed system serves as a platform for collaboration among librarians, digital humanities scholars, poetry experts, and AI researchers. The collaborators' active participation in every step of the workflow increases the quantity and quality of the descriptive metadata, and mitigates the potential bias caused by large-scale AI models. The workflow also involves a feedback routine that first collects evaluation results from librarians and general users, and then reinforces the system's performance and fairness. In the following subsections, we will see how each component in the workflow is defined and who contributes to which modules in which roles. The project work plan spans the entire duration of the project as shown in Table 1.

Year	Project Task	Outcomes
1	Data Collection	● Poetry collections ○ Underrepresented lyric poetry ○ General lyric poetry ● Auxiliary data collections ○ Author's notes, biography ○ Users' and critics' commentary ● HTRC worksets of the poetry and auxiliary data
	Annotation	● Emotion and theme metadata of 1,000 poems ● Annotation software
2	Modeling Building	● Emotion classification AI models ● Theme classification AI models
	System development	● Emotion and theme-based poetry discovery system for librarians with HTRC Analytics ● Emotion and theme-based poetry discovery website for the public with public domain poetry
	Toolkit development	● Open toolkit for librarians and poetry organizations
3	Evaluation and Update	● Interviews with librarians ● User surveys
	Toolkit refinement and publish	● Refine and publish the toolkit

Table 1. The timeline of the project work plan.

2.4. Data Collection

Building a labeled dataset is an essential part of training an AI model, although the dataset is often prone to bias due to the complexity and cost of the task. The project team proposes a careful data collection mechanism that overcomes potential annotation inaccuracies, inconsistencies, and biases, by focusing on the demographics of the poets, variety of themes, and balance between different social groups. In addition, we propose employing various auxiliary data that indirectly describe the poem's characteristics, turning the AI system into a multi-modal analysis tool.

Poetry collections

The project team will identify any historically underrepresented poet groups regarding their gender, race, and other social groupings with the help of the advisory board. We will then build an underrepresented poetry collection from English and American lyric poetry from the 19th to the 21st century. In addition, we will also build a general poetry collection, which will serve as a control group. By contrasting the two collections, we will examine how diversity and inclusion in the data collection mitigate biases of AI models. To make the collection visible to librarians and digital scholars, we will build matching HTRC worksets. The HTRC workset is a direct input source for any HTRC algorithms, which can be assessed by any librarians and scholars through HTRC (Jett et al., 2016).

Detailed analysis of these poetry collections not only increases the usability of the data, but helps identify and disclose any potential bias within those sets. To this end, the project team will seek guidance from the librarians, poetry experts, and the IU Luddy Assistant Dean for diversity and inclusivity on the advisory board to list known underrepresented demographic groups. The advisory board will recommend various resources of such poetry, such as bibliographies, anthologies, and online resources about poems; e.g., Columbia Granger's Index to African-American Poetry (Frankovich & Larzelere, 1999), an anthology of modern American poetry (Nelson, 2000), and Asian American Voices in Poetry in poetryfoundation.com, a librarian's resource curated by a well-managed poetry website. Last but not least, the project team will categorize poems into two groups, the public and copyright-protected domains, so that direct retrieval of the poem text using our system is mindful of copyright issues.

Auxiliary data collections

The project team proposes to use auxiliary data to supplement the hard-to-interpret poetry text, such as the author's notes and commentaries made by the readers and critics. For example, "The Road Not Taken," by Robert Frost, is widely interpreted as an ode to individualism; however, Frost's notes imply an opposite interpretation[1]. From both the AI model and human reader's perspective, auxiliary data is easier to interpret, as it is usually written in more straightforward language. The PD has previously studied similar ideas regarding song lyrics. She discovered that music listeners' comments could provide additional clues to aid an AI model's attempts to understand themes from song lyrics (Choi et al., 2016). Due to the similar characteristics between poetry and song lyrics, the extended use of auxiliary data, such as the critics' commentaries, user interpretations, authors' notes, and biographies, could also be applied to the proposed unbiased poetry analysis system. To begin with, the project team will collect the data from two sources: poetry websites (Poets.org) and poetry commentary books. Once again, we will be careful with the copyright of the comments and auxiliary data.

2.5. Data Annotation

Based on the careful analysis of the available poetry collections, the project team will build a few collections that consist of underrepresented groups of poems. Then, we will collect the emotion and theme labels of those collections to train AI models. The PD will apply her previous experiences designing, developing, and managing annotation systems used by diverse user groups for annotating K-pop songs (Lee et al., 2013; Hu et al., 2014) to the annotation study. Furthermore, to get high-quality labels, we will recruit annotators from a carefully chosen group of readers; i.e., English majors at Indiana University, rather than relying on crowdsourcing frameworks. To ensure we recruit annotators from various demographics, we will advertise the positions to groups of underrepresented populations (details are included in the Diversity Plan). Nikki Skillman, a faculty member in the English department at IU and on the advisory board, will oversee the project team's design and analysis of the annotation study. We plan to annotate at least 1,000 poems, each of which is annotated by at least three participants.

The annotation user study will collect two types of metadata—themes and emotions. Themes can exist across multiple layers and are often subjective, as readers' own experiences determine them. In this study, we will collect a single-word theme as a starting point. As for the emotion labels, we will use both emotion adjectives and a two-dimensional emotion space defined by *arousal* and *valence* as defined in psychology (Russell, 1980). These representations are most commonly used in sentiment analysis research, although they have limitations due to their simplicity (Kim et al., 2010). The annotators will use either a graphical user interface to mark their perceived feelings about the poem in the two-dimensional emotional space or to type emotional adjectives.

The annotation system will be developed and hosted on a computer without an Internet connection to protect copyright-sensitive data. We expect that each annotator will annotate five poems per hour, although we will adjust the

[1] https://www.theatlantic.com/video/index/555959/robert-frost-road-not-taken/

amount of work and the compensation for the test participants accordingly. The Indiana University IRB will examine the annotation study. The collected metadata will be published through Indiana University DataCORE, a repository for sharing and archiving research data, developed at Indiana University.

2.6. Model Design and Building

The project team will actively employ recent deep learning-based advancements in NLP and text mining as a starting point to develop AI-based poetry analysis systems. We also propose new methodologies to improve these baseline methods because the unbiased poetry analysis model will consider various aspects of data collection, annotation, and computational resources.

The basic transfer learning pipeline

Similar to other AI model training tasks, our training process will inevitably suffer from a lack of labeled data. The project's rigorous data annotation effort is designed to help mitigate this issue, but the amount we collect is still far from enough to train a large deep learning model. A common practice in deep learning is to *transfer* the model trained from a similar task to the target task (Weiss et al., 2016). In our case, we can utilize state-of-the-art text processing models, such as BERT and GPT-3, that have been trained from a massive amount of text data. However, the original task used to train these baseline models is different from emotion or theme-based classification of poems (e.g., predicting a missing word from a sentence). Transfer learning allows us to capitalize on some of the model's abilities that successfully recognize the basic meanings of the words within their context, assuming that they are universally useful for any text-related applications. Then, we will use our small labeled set of poetry data to *fine-tune* the model to fill in the gap between what the initial BERT model can do and what we want to achieve in our poetry analysis tasks.

Emotion classification

It is rare to find poems annotated according to the emotion they convey. Hence, our annotation study will produce a useful new dataset from which our models are fine-tuned. In addition, as part of the transfer learning process, the team will explore the possibility of leveraging similar kinds of text collections, such as song lyrics. Because song lyrics have been better annotated with emotion-related tags than poems, and are also available in public datasets (Delbouys et al., 2018), a model learned from this similarly difficult text can be transferred to the proposed poetry analysis tasks with small adjustments. Also, as for emotion analysis, we believe that incorporating the well-studied word-level affective scores into the deep learning model as a *conditioning mechanism* is a sensible approach (Dumoulin et al., 2018).

Theme classification

We will develop the theme classification models using the transfer learning pipeline as well: initializing the classifier with the pre-trained BERT model's parameters and then fine-tuning it via our annotation data. In addition to the basic pipeline, theme classification can also benefit from side information. In previous research, the PD discovered that people's interpretations of song lyrics could provide additional clues, because those commentaries are more plain-spoken and descriptive. In this project, the team will extend this idea and examine the potential of a multi-modal poetry analysis system using auxiliary data, such as the author's explanation of their poems, the poet's biography, and the readers' and critics' commentaries.

2.7. Initial Deployment, Evaluation, and Feedback to the Next Cycle

The final goal of the project is to publish a toolkit that consists of various open-source software components and documentation for librarians: (a) a tutorial on how to build unbiased HTRC worksets along with annotation

software; (b) web-based poetry browsing and searching interfaces via advanced metadata; (c) a backbone AI engine that performs metadata analysis on the poetry; (d) a tutorial on how to use the poetry discovery system in HTRC Analytics; (e) detailed documentation of all decisions at every stage of AI lifecycle along with video tutorials. The main purpose of this toolkit is to assist any librarians in building ethical AI models with their own collections.

At the end of the second year, as an initial evaluation mechanism, the system will be deployed to the main stakeholders, the digital librarians, and patrons, who will have early access to the initial version. For this initial deployment, we will repurpose the AI-based poetry analysis model as a part of HTRC Analytics so that various digital librarians in the HTRC community can use them in a standardized manner. Then, we will collect feedback on the open toolkit and the AI models by interviewing ten librarians whose areas are in Digital Scholarship, English and American Literature, or Library Technologies. As stated in the Diversity Plan, we will commit to recruiting at least 50% of librarian participants from historically underrepresented groups in terms of the demographics. Each librarian will be asked to use the poem discovery system with three different AI models regarding the level of biases in the HTRC platform following the guidelines in the open toolkit. Subsequently, the PD and a graduate student will conduct interviews with individual librarians to understand the system's usability, the effectiveness of bias mitigation approaches, and the usefulness of the open toolkit. After data analysis, including statistical analysis of interview responses, the project team will meet with the advisory board to discuss the findings and improve the technical part of the toolkit.

The project team expects this process to be iterative for a continuous and sustained improvement of the system. The project team's goal is to accommodate various real-world use cases, such that the participating librarians can devise various new curation options that highlight underrepresented author groups and specific themes that their local patrons might find interesting. In this human-in-the-loop process, the librarians' feedback will be actively utilized to improve the efficacy of the entire AI model training workflow. Hence, in the third year, the project team will re-adjust the entire workflow—the data collection will be re-evaluated to include more poets for better diversity, annotation results will be re-weighted to take into account any inaccuracies caused by the labeling process, and finally, a re-training of the AI model will follow accordingly. The project team will also collaborate with our advisory board member, Jon Dunn, who is directing the AMP metadata-generation ML tool development project. His expertise in developing the AI-based metadata system for audiovisual content and interfacing with the librarians using that system will be of great help.

In the final year, the project team will use the web interface to perform a user study as well. To that end, the project team will conduct surveys with 50 general public users to understand their user experience with the system and the effectiveness of the different AI models in terms of the level of biases. We will commit to ensuring the diversity of the participants through the recruitment approach mentioned in the Diversity Plan section. The PD will adopt a user experience survey similar to the one for the evaluation of music discovery systems in her previous study (Hu et al., 2016). The website will also provide a platform for patrons who would like to discuss poems' themes and emotions with others; we will observe the patrons' participation patterns and apply our findings towards re-defining our workflows. The user studies will also go through Indiana University's IRB approval process.

2.8. Dissemination Plan

The project team strongly believes that our society benefits the most from the active and wide dissemination of the project results. First, the PD and a graduate student will present and publish research findings at major conferences and journals in library and information science, such as the Digital Library Federation Forum, ALISE, Digital Humanities, ASIS&T, iConference, ALA Annual Conference, JCDL, and JASIST. In addition, we will release detailed documentation in the form of a white paper describing in-depth technical details and guidelines.

A dedicated website will host a web-based interface for librarians and patrons across the nation. They can freely test out the advanced metadata-based browsing and searching functionalities. All source codes will be open-sourced at Github.com[2] for the public to follow up with more advanced AI models and metadata utilization. The project team will advertise the project results to various library listservs and social media to bring the community's attention to the project results.

2.9. Personnel

The PD, Dr. Kahyun Choi, will lead every stage of the project, including data collection, annotation, AI model building, system development, evaluation and fine-tuning of the models, toolkit development, and dissemination. In addition, the PD will communicate with the advisory board and guide students to integrate their work into a team effort. She has a diverse background, such as her Ph.D. study in Library and Information Science at UIUC, a software engineer position at a search engine company, AI-based music and lyrics analysis, and media art. Such an interdisciplinary background brings her capability to navigate the two worlds: AI and Libraries.

A Ph.D. student and an hourly graduate student will participate in the project. The doctoral student will participate in the project research and dissemination under the mentorship of the PD, and an hourly graduate student will be employed to program the web-based interfaces.

Advisory Board is composed of 5 people with expertise in poetry, library and information science, digital scholarship librarianship, metadata generation, ML in libraries, large-scale digital libraries, and diversity and inclusion. **John Walsh**, the Director of the HTRC and Associate Professor of Information and Library Science at IU, will bring his research experience with HTRC and his expertise in poetry. He will also introduce the team's effort to the HTRC community. **Jon Dunn**, Assistant Dean for Library Technologies at IU, has been leading multiple projects regarding building open-source metadata extraction software for audio-visual collections and deploying them in multiple libraries (IMLS grant number: LG-70-17-0042-17). **Angela Courtney**, Librarian for English & American Literature at IU, will help curate diverse collections, giving feedback on the toolkit from a librarian's perspective. **Nikki Skillman**, Associate Professor of English at IU, is an expert of modern and contemporary poetry in English. She will advise on the team's activities on the diverse poem collection building and annotation process. **Lamara Warren**, the Assistant Dean for Diversity and Inclusion in the Luddy School of Informatics, Computing, and Engineering at IU, will provide expert guidance on areas of diversity and inclusivity throughout the planning, development, assessment, and dissemination of the project. The research team will meet the advisory board members two to four times per year to get feedback.

3. Diversity Plan

The project's primary goal is to improve the fairness and unbiasedness of AI models. Also, the PD is committed to increasing diversity and inclusion in her work. For instance, the PD is currently serving as a mentor of the WiMIR mentoring program, whose goal is to increase opportunities for underrepresented groups in the field of Music Information Retrieval. Based on her experience as a first-generation college student, a female student in male-dominant disciplines, and an Asian immigrant, she understands how important it is to take action to increase diversity, inclusion, and equity in her professional community.

The project team will ensure that balanced demographics are represented as essential for diversity, equity, and inclusion when organizing the project team, the advisory board, annotators, and the focused group of patrons. The project team will advertise the annotator and user study participation opportunities to underrepresented populations within the IU community through various IU affiliated organizations for diversity and inclusion, such as the Asian

[2] https://github.com

Culture Center, First Nations Educational and Cultural Center, Neal-Marshall Black Culture Center, La Casa Latino Cultural Center, LGBTQ+ Culture Center, and Center of Excellence for Women & Technology. The PD will advertise the Ph.D. position to ALA Spectrum Scholarship awardees and various student organizations of underrepresented students in universities, similar to the centers at IU. When recruiting librarians for the interviews, we will advertise the opportunities to ALA-Affiliated Associations of Ethnic Librarian groups, such as the American Indian Library Association, Asian/Pacific American Librarians Association, the Black Caucus of the American Library Association (BCALA), the National Association to Promote Library and Information Services to Latinos and the Spanish-speaking, and Rainbow Round Table (RRT) of the American Library Association.

4. Project Results

The project aims at improving the readiness of the digital libraries for the adaptation to the rapidly evolving AI-based tools. Our project results are expected to be different from an ordinary black-box AI model trained blindly from a big dataset, which tends to be biased unfavorably to the underrepresented groups. Consequently, the project results can benefit all stakeholders of the digital libraries. First, considering the public roles that libraries often play, serving the local communities and marginalized social groups is of utmost importance. The project results will provide a guided framework to the librarians so they can actively participate in the process of customizing a relatable AI model for the community they are serving, strengthening the library-patron relationships, improving the overall fairness of AI models, and influencing other areas. Second, the general public can also benefit from the project results as they can explore the massive digital collection of poetry using richer metadata tags. The enhanced interface between the patrons and digital libraries can also increase the exposure of underrepresented entries, leading to increased diversity in the literature. Since the proposed framework activates readers' participation by encouraging them to share their own interpretation and understanding, the patron's role can evolve into a more creative one instead of a mere consumer. Finally, digital humanities scholars can also benefit from the fairer, thus more accurate, AI models to conduct a large-scale analysis on literary works, such as trend analysis, the discovery of author networks, etc.

The project results will be open-sourced under the Creative Commons Attribution 3.0 License to maximize sustainability. All the source codes of web-based interfaces and trained AI models will be published via Github.com so that interested researchers can "fork" and improve over time. The documentation on the guided workset builder for the integration with HTRC Analytics can also encourage the future versions of the project results to be utilized as a part of the HTDL framework. In that way, the concepts and framework developed during the project period can reach out to the larger communities even after the end of the project.

			SE	OC	NV	DE	JA	FE	MR	AP	MY	JN	JL	AG
Year	Broad Tasks	Activities												
Year 1	Data Collection	Build poetry collections	■	■										
		Build auxiliary data collections			■	■								
		Build HTRC worksets of the poetry and auxiliary data					■							
	Annotation	Select poems for annotation						■						
		Request IRB approval for annotation study					■	■						
		Build annotation software					■	■						
		Recruit annotators						■	■	■	■			
		Conduct annotation study								■	■	■		
		Postprocess the annotated data											■	■
	Meeting with AB	Get feedback about data collection and annotation from individual AB members	■			■			■			■		
	Dissemination	Dissemination of initial research findings								■	■	■	■	■
	Model building	Set up a Deep Learning development infrastructure for model building in Year 2											■	■
Year 2		Build emotion classification AI models	■	■	■									
		Build theme classification AI models				■	■	■						
	System development	Build poetry discovery system for librarians as part of HTRC Analytics							■	■				
		Build poetry discovery website for the public with public domain poetry							■	■				
	Toolkit development	Build an open toolkit for librarians and poetry organizations									■	■	■	■
	Meeting with AB	Get feedback about the models, systems, and open toolkit from individual AB members	■			■			■			■		
	Dissemination	Dissemination of interim research findings								■	■	■	■	■
	Evaluation and Update	Request IRB approval for the evaluation studies in Year 3									■	■		
		Recruit librarians and participants for user studies in Year 3											■	■
Year 3		Conduct interviews with librarians and user surveys with general public users	■	■	■									
		Data Analysis				■								
		Update the ML pipeline based on findings from the user studies					■	■						
	Toolkit development	Refine the toolkit							■	■	■			
		Publish the toolkit									■	■	■	■
	Meeting with AB	Get feedback about evaluation and the toolkit from individual AB members	■			■			■			■		
	Dissemination	Dissemination of final research findings								■	■	■	■	■

Digital Products Plan

Poetry Metadata

Type: Our annotation study will produce theme and emotion labels for about a thousand targeted poems. They will be represented in a standardized format such as json or Dublin Core metadata. Each file will also contain other identifying information, including author names, title, DOI, and poetry text itself when it comes to public domain data.

Availability: The team will use the Indiana University Scholarly Data Archive (SDA; https://pti.iu.edu/storage/sda) to store our data safely and to allow easy access for the public. It is a distributed storage service via mirrored tape silos in Bloomington and Indianapolis, IN. The data will be accessible through the IU DataCORE repository, and they will be listed and hosted on the dedicated project website that the project team will build. Librarians, scholars, and the general public will find the data through the website.

Access: The poetry text itself will be in the public domain with no copyright or proprietary restrictions, and the metadata generated as a result of the project will be open to public usage via Creative Commons Attribution 3.0 License. By doing so, anyone is free to copy and redistribute our metadata as long as they give appropriate credit. The librarians can also build upon our initial metadata collection by explicitly stating the changes they make. We have chosen this type of license because we want many librarians and future researchers to use our dataset to build their own workset based on their needs and improve the main AI module in the future.

Sustainability: For sustained use of the collected annotation data during the project period and beyond, the project team will work with the IU Libraries and University Information Technology Services (UITS) to make the resources easily accessible by the public. The PD will make sure that the data is available for at least three years beyond the termination of the project period via the combination of the IU SDA and IU DataCORE repository.

Source Codes

Type: One of the main technical activities of the project is to develop AI models that analyze the poetry text and estimate its theme and emotion labels. Training such a model can be done via one of the widely used Python programming packages specialized in machine learning system development, such as TensorFlow and PyTorch. The team will produce various software modules written in these programming languages, which others can reuse to reproduce and improve the models.

Availability: In addition to the combination of the IU SDA and IU DataCORE repository, the team will create a group account in the widely used software dissemination service, github.com. The group account will host a few source code repositories to distinguish different models from each other. The github.com repositories will be accessible free of charge to the public and can create an ecosystem by monitoring any branching-out projects. It is a common custom in the research community to manage open source libraries via github.com.

Access: The source codes will be distributed with the Creative Commons Attribution 3.0 License. Once again, it will allow free access and reproduction of the code by anyone; it will also allow changes to be made to the base code with proper documentation of changes. In this way, we encourage any subsequent work to improve the performance of our models via a community effort.

Sustainability: We will keep the GitHub repositories open and accessible during the project period and at least three years after the end of the project. For better sustainability, given the third-party nature of the github.com service, the PD will collaborate with UITS, so the codebase is hosted and available for a prolonged period of time (3+ years after the project). We will use several archiving and hosting services within IU, such as IU SDA, a dedicated project website hosted by the IU Sitehost Service, etc.

Documentation, White Papers, and Research Papers

Type: The project will produce a detailed manual for the librarians to build their own workset within the HathiTrust Research Center (HTRC) framework. HTRC is a collaborative research center that facilitates non-profit and educational uses of the HathiTrust Digital Library by enabling computational analysis of works from its collection. Other research results on the user studies and AI models will also be published in the form of white papers and research papers. We will use PDF format for the digital distribution of this type of material.

Availability: The documentation, white papers, and research papers will be hosted and available to the public in various virtual locations. First, we will use IU DataCORE as the main permanent place to archive the documents. In addition,

whether they are peer-reviewed or not, we will also deposit the papers on arxiv.org, a popular archiving service used in various academic disciplines. We will also list up the documents and research papers in the dedicated project website hosted by IU Sitehost for public access.

Access: All the non-archival documents will be distributed to the public under the Creative Commons Attribution 3.0 License. Some peer-reviewed papers available through the publisher's websites might be with only limited access or can incur fees. In such a case, the project team will make an "author's version" freely available through other virtual locations, which are usually allowed by the publishers for personal use.

Sustainability: The project team will make sure that any publication is available during the project period and after the end of the project activity, at least for three years. Our effort to maintain the sustainability for these types of digital products will be similar to the above mentioned cases.

Data Management Plan

This plan describes the management, dissemination, retention, and archiving of the research data produced during the proposed project. The staff of the Luddy School of Informatics and Computing, with the assistance of IU Bloomington Libraries and University Information Technology Services (UITS), will provide for sustainable discovery of, access to, and preservation of these data for use by other researchers, instructors, and interested members of the public for the length of this project and at least three years beyond. This will be facilitated through data and publication deposits in existing open-access disciplinary and/or institutional repositories.

Roles and Responsibilities

The PD has primary responsibility for the collection, management, custody, and retention of research data. The PD will work with a graduate student who will assist with data collection, data management, and data analysis. When appropriate, the PD will delegate dissemination and preservation responsibilities to Indiana University Libraries.

Types of Data

The proposed project will generate metadata and user feedback through user studies, where surveys using standardized questionnaires and interviews will be the main methods to collect the data. In doing so, any identifiable information of participants will be removed from that data for privacy. The project team will also collect existing data that are currently publicly available from the web using Application Programming Interfaces (APIs).

We will hire at least three English major undergraduate students for the annotation effort. They will annotate poems with high-level metadata, including theme and emotion, that requires attentive reading. The collected metadata values will be stored along with other descriptive metadata to distinguish one poem from others. It will be approximately 10MB of data (one thousand 100KB files).

Interviews will collect feedback from ten librarians and 50 participating general users. Although we do not plan to publish the interview results directly, their recordings and transcriptions will be stored after anonymization.

Sensitive and confidential data collected, such as the original interview recordings will be treated following Institutional Review Board (IRB) regulations, and an added layer of security will be implemented using the separation of identifiable data.

Data Storage, Preservation, And Sharing

We will anonymize and remove personal information from the intermediate research data, such as the recordings of the interviews and survey responses. This kind of raw data will be shared among the project team via IU-affiliated cloud storage, such as Microsoft OneDrive and Google Drive, but not with the public. These cloud storage services are associated with the IU's CAS authentication system that supports two-factor login for increased security. This type of data needs provisions for confidentiality due to ethical restrictions and the protection of privacy. This sensitive data is governed by an IRB policy.

For the public portion of data, we will use IU Scholarly Data Archive (SDA; https://pti.iu.edu/storage/sda), a distributed storage service that is centrally supported across mirrored tape silos in Bloomington and Indianapolis. Data stored on the SDA will be made freely available and archived in the IU DataCORE repository, which will provide a user-friendly interface for the organization, context, and discoverability of data. This combination of IU DataCORE and the SDA provides mirroring, redundancy, media migration, access control, file integrity validation, embargoes, and other security-based services that ensure the data are appropriately archived for the life of the project and beyond the project if necessary.

In addition, the project team will maintain a dedicated project homepage through IU Sitehost web hosting service (https://kb.iu.edu/d/axnv) , which will also list locations of the major dataset and other documents.

Results, data, presentations, and videos will be made available to other researchers and the wider population on a timely basis. The data will be released to other researchers on IU DataCORE within two years of data collection.

Period of Data Retention

The PD will retain the collected data for ten years to allow the PD and her collaborators to analyze the data and publish the findings.

Restrictions on Data or Product Storage, Access, Preservation, or Sharing

To increase access to the published research that has been funded, the research collaborators will deposit peer-reviewed or pre-print manuscripts (with linked supporting data where possible) in the IU DataCORE institutional repository. In addition, the project team will also utilize other third-party archival services, such as arxiv.org, for better accessibility. The documents will be also listed and hosted on the dedicated project website.

Source Codes

Source codes are an important part of the project results. In addition to the usual Python source codes, we also plan to disseminate fully trained AI models, so the librarians and the public can still use the AI system without recompiling and rerunning the code to reproduce the models. It is a critical contribution to the community, as training those models requires expensive computing facilities and professional efforts. We will use github.com to archive the source codes and make them available to the public with a minimal copyright claim: Creative Commons Attribution 3, which allows free access and manipulation.

Data Formats

The following data formats will be used: text documents (PDF), metadata files (e.g., json), websites, and raw source codes and binary model checkpoints (e.g., py, ipynb, pt, etc.). We will utilize the Dublin Core metadata scheme to capture information about the data collected during the course of our research. We will work with a metadata expert from IU Libraries to create a working template that captures each dataset's metadata throughout the research process. Upon completion, we will export these data to Dublin Core format, which conforms to the data submission requirements of the IU DataCORE and many other relevant museums/repositories.

Plans for Archiving and Preservation

The IU Libraries and University Information Technology Services (UITS) will assist the PD to provide for sustainable discovery, access to, and preservation of these data for use by other researchers, instructors, and interested members of the public for the length of this project and at least three years beyond. When appropriate, the PD will delegate dissemination and preservation responsibilities to Indiana University Libraries.

Organizational Profile

Our mission

The mission of the Luddy School of Informatics, Computing, and Engineering at Indiana University Bloomington is to excel and lead in education, research, and outreach spanning and integrating the full breadth of computing and information technology, including the scientific and technical core, a broad range of applications, and human and societal issues and implications. The vision of the School is of a community committed to diversity as a core strength and as a principle for maximum innovation, creativity, pedagogy, and scholarship.

Governance structure

The Luddy School of Informatics, Computing, and Engineering at Indiana University Bloomington is currently led by interim Dean Dennis Groth, who reports to the Indiana University Bloomington Provost and Executive Vice President, Rahul Shrivastav. Provost Shrivastav reports to IU's President Pamela Whitten, who then reports to the IU Trustees. IU Trustees are the governing body of Indiana University; for a full list of current trustees please visit this link: https://trustees.iu.edu/the-trustees/current-trustees/index.html.

Service Area

Luddy School of Informatics, Computing, and Engineering is located at the center of the Indiana University Bloomington campus, with enrollment over 45,000 students in fall 2021. Located in southern Indiana about an hour from Indianapolis, the surrounding region is rural, where 24.3% of households fall below the poverty level, nearly 10% higher than the state average. Indiana University also encompasses satellite campuses, including Indiana University-Purdue University Indianapolis, Indiana University East, Indiana University Kokomo, IU Northwest, IU South Bend, and IU Southeast where total enrollment over 91,000 students. Both the academic community and general public are the central focus/main beneficiaries of our research efforts, largely through our partnership with Indiana University Libraries. All IU libraries are open to residents of the state as well as to Indiana University faculty and students. A team of specialists select, manage, and build the library research collections, which include more than 11,532,115 million cataloged items. The materials support every academic discipline on campus, with an emphasis in the humanities and social sciences, for instance at the Herman B Wells Library. Collections also include journals, maps, films, and sound recordings. Users can access more than 1,871 databases, 60,000 electronic journal titles, and 1.9 million electronic books, as well as locally developed digital content. IU Libraries is prolific in open access publishing and hosts 40 open access journals. A general audience will also benefit from this project as all digital resources will continue to be accessible and discoverable online.

History

The School of Informatics was established in 2000 as the first of its kind in the US. In 2005, the Department of Computer Science joined the School, significantly advancing the program. In 2013, the School of Informatics merged with the School of Library and Information Science and became the School of Informatics and Computing. In August 2017, the name of the School of Informatics and Computing on the Bloomington campus was officially changed to the School of Informatics, Computing, and Engineering (SICE), and in 2019 the Bloomington school was renamed Luddy School of Informatics, Computing and Engineering in honor of Indiana University alumnus and information technology pioneer Fred Luddy. Luddy Hall opened in 2018, and houses three departments—computer science, information and library science, intelligent systems engineering—plus undergraduate and graduate advising, career services, administration offices, and several high-tech classrooms, labs, and makerspaces.