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

An analysis of poet demographic and thematic diversity in a poetry collection for inclusive AI

Kahyun Choi and Gyuri Kang

DOI: https://doi.org/10.47989/ir30iConf47263

Abstract

Introduction. AI technologies, such as theme classification and named entity recognition, enhance digital library accessibility. However, they may introduce biases if training datasets lack adequate representation. For instance, prior AI models for poetry classification overlooked dataset diversity, raising concerns about representation. To address this issue, this study assesses the dataset representation and examines potential issues in AI model design for poetry collections.

Method. We annotated and published the race and ethnicity of poets in an American poetry collection curated by *poets.org*, which was recently used to train a poetry theme classification system. We then examined the diversity of the collection using these annotations.

Analysis. We compared the racial/ethnic composition of the collection to U.S. Census data and conducted group-exclusive top word analysis, popular theme analysis, and entropy-based analysis of theme distribution diversity to evaluate linguistic and thematic diversity.

Results. Our findings indicate that most underrepresented groups are well-represented in the collection, except for Latino/a/x American poets. Furthermore, we found that poems from underrepresented groups increase the collection's linguistic and thematic diversity.

Conclusions. To design responsible AI that embraces diversity, it is essential to assess dataset representation and support non-standard English and diverse themes beyond those popular with the general population.

Introduction

Artificial intelligence (AI) has unlocked the potential for enhanced recommendation and search services within online library collections. However, careless use of AI can introduce pitfalls, such as biases embedded in the collections upon which AI models are based, leading to the models' biased outcomes (Cordell, 2020). Indeed, collections often do not represent the general population accurately, causing models to discriminate against marginalized groups (D'ignazio & Klein, 2023). For instance, the recent large language model, GPT-3, demonstrated biases against Muslims and other marginalized groups (Bommasani et al., 2021). Thus, investing extra effort and attention in data collection for AI models, with an emphasis on equity and representation, is crucial (Jo & Gebru, 2020). Moreover, it is critical to assess the equity and diversity of pre-existing collections before use.

Addressing this challenge becomes urgent in the poetry domain as readership rises and the application of AI in poetry analysis expands. National endowment for the arts survey has revealed a 76% increase in US poetry readership from 2012 to 2017 (Iyengar et al., 2018). A follow-up survey in 2022 shows a similarly high engagement, as 11.5% of US adults engage with poetry through reading or listening (Iyengar, 2023). Alongside, various AI systems for improving accessibility of poetry collections, such as poetry theme classification have been developed (Rakshit et al., 2015; Lou et al., 2015; Kaur & Saini, 2017; Navarro-Colorado, 2018; Choi, 2023). Yet, these studies have not examined if the collections were diverse enough to accurately represent the general population. Thus, it remains uncertain whether they adequately account for poems by poets from underrepresented groups.

As part of the '*unbiased AI for poetry analysis: toward equitable and diverse digital libraries*' project funded by institute of museum and library services (IMLS), our study addresses this oversight by assessing *poets.org*'s curated poem collection, with a focus on the race and ethnicity of the poets. We selected this collection because it was used to train a theme classification system in one of the most recent AI systems for poetry (Choi, 2023). Specifically, we compared the U.S. Census population data on race and ethnicity with those of our poem collection to determine if the collection accurately reflects the general population. Furthermore, we analysed prevalent words and themes in poems written by these groups. We investigated how underrepresented groups' work contribute to the word and theme diversity of the poem collection. Also, we identified potential issues with AI systems that are developed mostly based on poems by dominant groups, while works from underrepresented groups are disregarded as outliers. While our primary focus is on this poetry collection, we suggest that our methodology for assessing demographic representativeness of a collection can be applied to other literary genres.

Diversity analysis

Data collection and pre-processing

We utilized the same collection of poems retrieved in October 2022 that Choi (2023) used to train their theme classification system. The poems are from *poets.org*, which is managed by the Academy of American Poets, the nonprofit charitable organization dedicated to fostering American poets and poetry. This collection features a wide variety of poems, each accompanied by several descriptive tags. We excluded audio-only entries, duplicates, and excessively short, resulting 9,445 works. As our project focuses on modern and contemporary American poetry, we selected 8,912 poems published after 1890, covering the era from Emily Dickson and Walt Whitman to 2022. The list of poets, categorized by race and ethnicity, can be available via this link: https://doi.org/10.6084/m9.figshare.25572459.v1

We have identified the racial and ethnic groups of poets using the '*occasions*' field in the `poets.org` collection, especially when the occasion is relevant to specific racial and ethnic groups.

To align with US Census racial and ethnic categories, we selected the following occasions: 'Asian/Pacific American Heritage Month' for Asian/Pacific Americans (APA), 'Black History Month' or African Americans (AA), 'Hispanic Heritage Month' for Latino/a/x American (LXA), and 'Native American Heritage Month' for Native Americans (NA). To better follow US Census categories, we distinguished between Asians and Pacific Islanders by manually identifying poems by Pacific Islanders through a review of the descriptions of the poems. Therefore, in this paper, APA-AA represents Asian Americans without Pacific Islanders, while APA-PA denotes Pacific Islanders exclusively. In our study, 'Others' or 'General' poems represent works by poets who are not associated with the specific occasions we use to identify the underrepresented groups. Additionally, we have reassigned 329 poets to specific underrepresented groups after reviewing their biography pages on `poets.org`; however, these biographies are not detailed enough to distinguish the mixed-race identities recognized by the US Census. Upon determining the racial and ethnic categories of the poets, we organized their poems into the corresponding categories. However, we acknowledge that our grouping strategy may overlook some poems by underrepresented groups, despite its comprehensiveness.

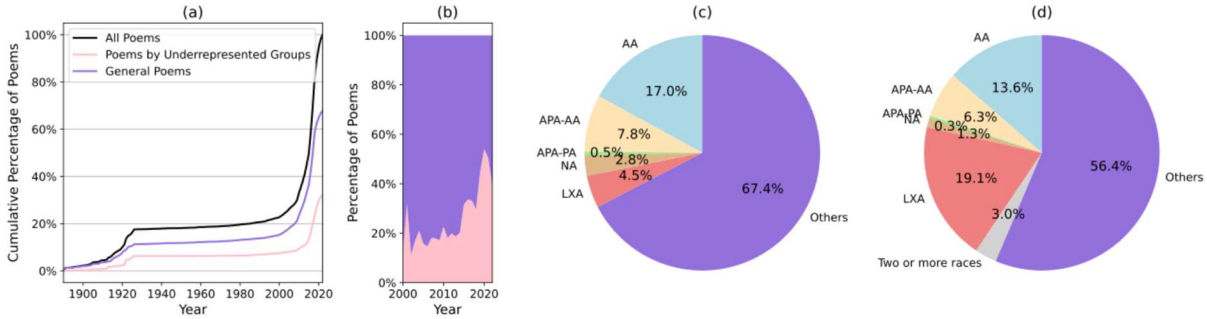

Figure 1. Representation and trends of poet groups in the United States: a) cumulative count of poems; b) yearly ratio of poems by underrepresented groups and general poets since 2000; c) composition of poems by group; d) U.S. population composition

Representation analysis of poems by groups

We examined the cumulative percentage of poems written by poets from underrepresented groups, by other poets, and both combined over time (see Figure1-a). The year indicates either the date of original publication or the date when *poets.org* made the work available on their website. In the collection, there is an increase until mid 1920s, followed by a plateau until early 2000s, and then a soaring upward trend. This pattern can be understood in the context of copyright laws; poems published before 1929 are copyright-free and can be published without permission, while those released after 1929 can only be published on the web with permission. Since its launch in 1996, `poets.org` has published approximately 80% of poems that are copyright-protected with permission. The scarcity of poems between the mid-1920s and early 2000s is attributed to the copyright issues or potential overriding of original publication year by the year of publication on `poets.org`. Over the last two decades, we can observe rapid growth in both general poems and those by underrepresented groups; notably, the latter have exhibited a sharp upward growth trend in the last decade, with recent years showing their proportion to be about half (see Figure 1-b), indicating *poets.org*'s commitment to equity and diversity.

Furthermore, we compare the composition of poems by individual groups with U.S. population demographics (U.S. Census Bureau, 2022) to evaluate how well poems from underrepresented groups are represented in the dataset. Figures 1-c and 1-d show that most underrepresented groups have higher representation, except for LXA: while the LXA population in the U.S. accounts for 19.1%, their representation in the collection is only 4.5%. This gap might be due to the prevalence of Spanish in the U.S.: 13% of Americans speak Spanish at home (Dietrich & Hernandez, 2022), and it remains the most popular second language (American Academy of Arts & Sciences,

2016). Thus, substantial portion of the American population may prefer reading poetry in Spanish. However, since English is the primary language of *poets.org*, most works are translated into English, which may affect the representation of LXA poets in the collection. Nonetheless, increasing the number of LXA poems, whether in the original language or translated, would further enhance the collection's already strong diversity and equity. This enhancement is also essential for reducing potential biases in AI models for poetry caused by data imbalance.

Groupwise exclusive top word analysis

Figure 2 shows the top 10 unique words exclusive to each group to explore the word diversity and cultural depth each contributes. AA poems feature colloquial language reflective of African American Vernacular English (AAVE) (Khera, 2021), with terms such as 'lawd,' 'hyeah,' 'souf,' 'lovah,' and 'whah,' which mean 'Lord,' 'hear,' 'south,' 'lover,' and 'where,' respectively. APA-AA poems are rich in cultural references and names common in Asian contexts, such as 'Shiratama' (a Japanese dessert), 'Chang' (a family name in Korea or China), 'lola' (grandmother in Tagalog), as well as names such as 'Acequia.' APA-PA poems prominently feature words from indigenous languages, including Hawaiian Pidgin (Roberts, 1995), such as 'nalani,' 'kai,' 'huki,' 'olelo,' meaning 'the heavens', 'sea', 'to pull,' 'language,' respectively, along with names of places they live, such as 'guam' and 'hawai.' LXA poems contain many Spanish words, including 'tata,' 'une,' 'alabanza,' and 'templo,' which translate to 'grandfather,' 'article,' 'praise,' and 'temple.' Finally, NA poems incorporate historically significant words that describe their tribes or towns, such as 'Spavinaw,' 'Shawnee,' and 'Anishinaabeg,' as well as terms like 'chieftain' and 'clans' to depict their societal organization. The inclusion of non-standard English and foreign words underscores the diversity that these groups bring to American poetry. However, this diversity also highlights the potential limitations of linguistic analysis tools, including named entity recognition when applied to American poetry, especially if they were not trained from non-standard English varieties such as Hawaiian Pidgin or African American Vernacular English, or on foreign terms and languages.

Figure 2. Word clusters of individual groups

Groupwise popular theme analysis

To examine the similarities and differences among groups, we first identified each group's top 10 most popular themes. To further assess the thematic diversity that each group contributes to the collection in comparison with general poems, we first subtracted the theme proportions found in the general poems from those in each specific group. We then ranked these differences and selected the top 10 themes that were most distinctively prevalent in each group. Table 1 presents these themes and their corresponding percentages: the numbers in parentheses in the first five columns represent the proportion of each theme within a group, while the numbers in the last four columns show the difference from the corresponding proportions in general poems.

The dominant themes in general poems such as nature, love, death, body, and existential, illustrate a broad range of human emotions and experiences. Each underrepresented group has its distinct set of most popular themes, which only partially overlap with those of general poems, indicating their unique thematic focuses. Specifically, the NA group has only two overlapping themes, LXA has three, while AA and APA each have six overlapping themes. Furthermore, we examined the relatively popular themes in comparison to the general poems to understand the unique thematic focuses within individual groups. Among all groups or three out of the four, themes such as America, ancestry, body, identity, and family are prevalent. History, immigration, migration, and

social justice are also prominent, appearing in two groups. These common themes among them reveal how their experiences as underrepresented groups in America impact their poetry. Unique themes present in only one group include beauty, death, and slavery in AA; fathers, mothers, and politics in APA; memories and violence in LXA; and earth, environment, landscapes, language, and nature in NA. These highlights unique cultural and historical characteristics of each group. Particularly, NA poetry stands out with the most distinctive themes, reflecting their unique history as the original inhabitants prior to European arrival, and their strong connection to nature and environment. The distinct sets of popular themes among underrepresented groups raise concerns regarding the current theme classification systems. As theme sets are typically selected based on average theme popularity, as shown in Choi (2023), themes that are popular only within underrepresented groups often get overlooked. To create more equitable AI systems for poetry, we suggest developing multiple sets of classifiers for underrepresented groups, not only generalized but also tailored to their unique themes.

	Top 10 Most Popular Themes					Top 10 Most Relatively Popular Themes			
	General	AA	APA	LXA	NA	AA	APA	LXA	NA
1	nature (9.6)	identity (15.2)	body (11.6)	body (14.3)	nature (22.1)	identity (11.0)	identity (6.8)	identity (9.4)	ancestry (15.9)
2	love (8.8)	body (12.3)	identity (10.9)	identity (13.6)	ancestry (17.8)	america (9.0)	family (6.0)	immigration (8.0)	nature (12.5)
3	death (8.4)	america (12.1)	family (9.3)	death (10.8)	body (13.4)	ancestry (7.6)	ancestry (5.9)	family (7.5)	environment (9.7)
4	body (7.9)	death (11.9)	death (9.0)	family (10.8)	landscapes (11.9)	history (5.6)	immigration (5.4)	america (7.0)	earth (8.6)
5	existential (7.6)	love (10.8)	nature (8.9)	america (10.1)	environment (11.9)	social justice (5.4)	america (4.1)	ancestry (6.4)	america (8.4)
6	self (6.2)	ancestry (9.4)	self (8.6)	immigration (8.5)	america (11.5)	body (4.4)	body (3.7)	body (6.4)	landscapes (7.7)
7	beauty (5.6)	self (9.1)	existential (7.7)	ancestry (8.3)	history (10.7)	death (3.5)	mothers (3.7)	social justice (4.6)	family (7.4)
8	animals (5.3)	nature (9.2)	ancestry (7.7)	memories (7.8)	family (10.7)	beauty (3.3)	migration (3.1)	violence (4.5)	language (7.3)
9	loss (5.2)	beauty (9.0)	america (7.2)	violence (7.3)	earth (10.7)	slavery (3.3)	politics (3.1)	memories (3.8)	history (7.2)
10	writing (4.5)	history (9.0)	animals (6.7)	loss (7.0)	language (10.3)	hope (3.1)	fathers (3.1)	migration (3.5)	body (5.5)

Table 1. Top 10 most popular themes per group and relatively popular themes per group against general

Entropy-based analysis of theme distribution diversity

We explore the theme diversity within each individual group of poets to assess their contribution to the collection's overall diversity. For this analysis, we use entropy, given that this has been widely used to assess the diversity of systems and environments based on the richness and evenness of values (Jost, 2006). Entropy is defined as $H = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$ where H is the entropy, n is the number of themes, $p(x_i)$ is the proportion of the i-th theme within the poem group, and the summation is across all themes A higher entropy suggests a greater number of associated themes per poem (richness) and/or a more equitable distribution among them (evenness).

Figure 3 displays the cumulative distribution functions (CDF) of all groups, which represents evenness of theme distribution. The x-axis represents the proportion of themes within each group, while the y-axis represents the cumulative probability. A curve that rises quickly to 1 indicates low evenness, as this suggests that a few themes dominate; the corresponding histogram would show a steep decline. Conversely, a curve that rises slowly and stops at a larger theme proportion indicates high evenness, suggesting a more uniform distribution of themes; the corresponding

histogram would decrease more gradually. Besides the CDF, the associated entropy scores and the average number of themes associated with each group, as a measure of richness, are presented in the legend of the figure. Because APA-PA associates with fewer than half of the themes, it cannot be compared with others fairly. So, we merged it back to the APA category instead of distinguishing it from Asian Americans in this graph.

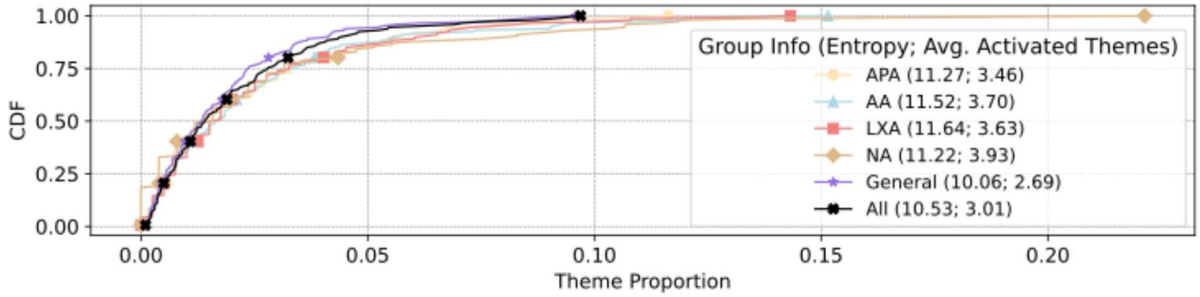

Figure 3. Cumulative distribution function (CDF) of theme proportions across groups, with corresponding entropy scores and average number of associated themes per group

The analysis of entropy scores shows a consistent trend: underrepresented groups achieve higher scores than their general counterparts, with scores of 10.06 for General, 11.27 for APA, and 11.64 for LXA. Furthermore, incorporating the underrepresented groups' poems into the general collection has increased the overall diversity by 0.47. To assess the contribution of richness to the entropy scores, we examined the average number of associated themes in each group. Compared to the General category's 2.69, poems from underrepresented groups tend to have more themes annotated, ranging between 0.77 and 1.24 more themes, which supports *poets.org*'s extra care toward these groups. While there is a correlation between entropy scores and thematic richness, this increase is also attributed to the evenness observed in the CDF in Figure 3. The CDFs of the underrepresented groups rise more gradually and stop at a larger theme proportion compared to the CDF for general poems. Similarly, although subtle, the CDF for all poems also ascends more slowly and stops at a slightly higher theme proportion, indicating an evenness contributed by underrepresented groups. Overall, our findings indicate that poems by underrepresented groups contribute to the collection's higher thematic diversity, benefiting from *poets.org*'s more thorough annotation of these poems and their more even distribution of theme proportions.

Conclusion

This paper evaluates *poets.org*'s poetry collection for racial and ethnic representation, addressing a gap in previous AI and NLP studies on poetry that overlooked the collection's assessment, potentially introducing biases against underrepresented groups. The collection generally reflects the demographics of the US population, with most categories having higher representation, except for LXA poems. The groupwise word and theme analyses show the diversity these groups bring to the collection, stemming from their culture and history. However, AI systems need to accommodate non-standard English and foreign terms to be more inclusive. Moreover, tailored theme sets for each group, rather than relying on those for the average, could ensure more equitable and inclusive AI systems for poetry. This study focuses on racial and ethnic diversity, one of many facets of diversity. In future work, we will further investigate other aspects, such as gender and disability.

Acknowledgments

This work was supported by RE-252382-OLS-22 from the institute of museum and library.

About the authors

Kahyun Choi is an Assistant Professor in the School of Information Sciences at the University of Illinois at Urbana-Champaign. She earned her Ph.D. from the School of Information Sciences at the University of Illinois at Urbana-Champaign. Kahyun Choi's research interests involve the application of computational methods and machine learning algorithms to various modalities, including audio and text. She can be contacted at kahyun@illinois.edu

Gyuri Kang is a PhD student in Information Science at Indiana University Bloomington. Her research interests include digital environmental humanities, cultural analytics, and NLP. She can be contacted at gyukang@iu.edu

References

American Academy of Arts & Sciences. (2016). The state of languages in the U.S.: A statistical portrait. Cambridge, MA: Commission on Language Learning, American Academy of Arts & Sciences.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258

Choi, K. (2023). Computational thematic analysis of poetry via bimodal large language models. Proceedings of the Association for Information Science and Technology, 60(1), 538-542. https://doi.org/10.1002/pra2.812

Cordell, R. (2020). Machine learning and libraries: a report on the state of the field. Library of Congress.

D'ignazio, C., & Klein, L. F. (2023). Data feminism. MIT press.

Dietrich, S., & Hernandez, E. (2022). Language use in the United States: 2019. American Community Survey Reports. Retrieved from https://www.census.gov/content/dam/Census/library/publications/2022/acs/acs-50.pdf

Drager, K. (2012). Pidgin and Hawai 'i English: an overview. International Journal of Language, Translation and Intercultural Communication, 1, 61-73.

Iyengar, S. (2023). New Survey Reports Size of Poetry's Audience, Streaming Included. Accessed: April 7, 2024.

Iyengar, S., Nichols, B., Shaffer, P.M., Menzer, M., Grantham, E., Santoro, H., Moyseowicz, A., & Hall, E. (2018). US trends in arts attendance and literary reading: 2002–2017.

Jo, E. S., & Gebru, T. (2020, January). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 306-316). https://doi.org/10.1145/3351095.3372829

Jost, L. (2006). Entropy and diversity. Oikos, 113(2), 363-375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Kaur, J., & Saini, J. R. (2017, February). Punjabi poetry classification: the test of 10 machine learning algorithms. In Proceedings of the 9th international conference on machine learning and computing (pp. 1-5). https://doi.org/10.1145/3055635.3056589

Khera, T. (2021). What Makes African American Vernacular English Distinct and Complex. Dictionary. com. Dictionary. com, February 21.

Lou, A., Inkpen, D., & Tanasescu, C. (2015). Multilabel subject-based classification of poetry. Nature, 2218, 30-7.

Navarro-Colorado, B. (2018). On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry. Frontiers in Digital Humanities, 5, 15. https://doi.org/10.3389/fdigh.2018.00015

Rakshit, G., Ghosh, A., Bhattacharyya, P., & Haffari, G. (2015, December). Automated analysis of Bangla poetry for classification and poet identification. In Proceedings of the 12th international conference on natural language processing (pp. 247-253).

Stevenson, D. (2021). Application of Shannon Entropy Metrics to Cultural Diversity and Language Evolution. Academia Letters, 2.

U.S. Census Bureau. (2022). Race and Hispanic origin. Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045222