Computational Thematic Analysis of Poetry via Bimodal Large Language Models

Kahyun, Choi Indiana University, USA | choika@iu.edu

ABSTRACT

This article proposes a multilabel poem topic classification algorithm utilizing large language models and auxiliary data to address the lack of diverse metadata in digital poetry libraries. The study examines the potential of context-dependent language models, specifically bidirectional encoder representations from transformers (BERT), for understanding poetic words and utilizing auxiliary data, such as author's notes, in supplementing poetry text. The experimental results demonstrate that the BERT-based model outperforms the traditional support vector machine-based model across all input types and datasets. We also show that incorporating notes as an additional input improves the performance of the poem-only model. Overall, the study suggests pretrained context-dependent language models and auxiliary data have potential to enhance the accessibility of various poems within collections. This research can eventually assist in promoting the discovery of underrepresented poems in digital libraries, even if they lack associated metadata, thus enhancing the understanding and appreciation of the literary form.

KEYWORDS

Digital libraries, multilabel classification, context-dependent language model, auxiliary data, computational poetry analysis

INTRODUCTION

Recently, poetry has gained significant interest, stemming from the public's heightened awareness due to high-profile events and the rediscovery of its therapeutic values during the COVID-19 pandemic. For instance, following Amanda Gorman's recitation of her poem at the 2020 U.S. presidential inauguration, poets.org reported a dramatic surge in traffic to their site, with a 250% increase in visitors compared to the same day in the previous year (Academy of American Poets, n.d.). Her powerful performance and inspiring message, presented amid a global pandemic, successfully captured widespread attention and admiration. Moreover, poetry websites experienced a substantial uptick in traffic during the pandemic, as poetry positively influenced mental health during such challenging times (Acim, 2021). Indeed, the therapeutic benefits of literature have long been recognized (Hynes, 2019). For example, inscribed in the ancient Library of Alexandria was the phrase "The place for a cure for the soul," and Aristotle's Poetics highlighted the restorative power of literature through catharsis. Bibliotherapy, the active application of literature for healing purposes, has been practiced for decades, with poetry being a primary genre (Harrower, 1972; Mazza, 2016).

Given the social and therapeutic virtue of poems, increasing their accessibility through diverse metadata can benefit the readers. However, digital poetry libraries often struggle with limited availability of metadata. Some services offer theme metadata for a limited number of poems, often excluding amateur works. The scarcity of advanced metadata types can be attributed to the lack of automated poem analysis tools, such as AI-driven natural language processing (NLP) algorithms, capable of achieving human-level comprehension of poetry. Poetry's complexity, characterized by its figurative language and multiple layers of meaning, makes it difficult for both humans and machines to interpret. Consequently, poetry has seldom been the focus of computational analysis, unlike prose genres, such as news articles and product reviews. Only a handful of studies have focused on computational poetry analysis (Kao & Jurafsky, 2012; Rakshit, Ghosh, Bhattacharyya, & Haffari, 2015; Lou, Inkpen, & Tanasescu, 2015; Kaur & Saini, 2017; Navarro-Colorado, 2018).

To bridge this gap, we propose a poem topic classification algorithm utilizing a pretrained large language model, building upon the work of Lou et al. (2015) and Choi (2021). Lou et al. (2015) introduced the multilabel poem topic classification task and explored its potential using support vector machines (SVM) on Poetry Foundation's poem data. We extend this study by examining the applicability of pretrained large language models, inspired by findings from another work on song lyrics (Choi, 2021). Choi's work focused on the song lyrics, which were assumed to consist of hard-to-interpret choices of words, making computational analysis difficult. Context-dependent language models, e.g., bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018), was successfully introduced to discern the underlying meanings of certain figurative words better than non-contextual baselines, such as support vector machines (SVM). In the proposed system, we are based on the similar conjecture that pretrained large language models possess a certain level of inference power capable of understanding the underlying meanings of complex text, including poems, which is the focus of this work.

86th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – 31, 2023 |London, United Kingdom.

In addition, to overcome the limited straightforwardness of the poetic text, we propose utilizing supplementary information, such as author's notes. From both the AI model and human reader's perspectives, certain types of auxiliary data are easier to interpret due to their more straightforward language. In previous research on song lyrics, for example, music listeners' comments successfully assisted AI models in understanding themes within song lyrics (Choi et al., 2016; Choi, 2021). In this work, we propose to use a different kind of auxiliary information, such as author's notes, by assuming that they can demystify symbolic text written by authors themselves. With this in mind, we focus on the following research questions:

Q1. Can context-dependent language models, such as BERT, decipher themes of poems more effectively than context-independent models, such as SVM?
Q2. Do authors' notes on their poems provide additional information to the poem topic classification system?

EXPERIMENT DESIGN

Data

In our study, we utilized data from poets.org, a nonprofit organization that offers online resources for poems, poet biographies, essays on poetry, and K-12 educational materials. Established in 1996, this pioneering platform has significantly contributed to the promotion and accessibility of poetry for a diverse range of audiences (Academy of American Poets, n.d.). For our research, we use the text of poems, their themes categorized by the platform, and supplementary information about the poems, such as author’s notes. The data were scraped in October 2022 using BeautifulSoup, a widely used Python library for parsing HTML documents (Richardson, 2007). Subsequently, data preprocessing removed non-ASCII characters and HTML codes to establish a clean, uniform dataset. After eliminating duplicates, we obtained a total of 11,803 poems. Among them, 3,581 poems include author explanations of their works in plain language, referred to as the auxiliary data in this study.

The auxiliary about-poem data has an average word count of approximately 18, significantly shorter than the poems themselves, which average around 207 words. Nevertheless, they tend to contain essential information about the poem. For instance, an excerpt from “Message to My Sistah” by Joe Balaz, “I just did // wat I talked about doing // wen we wuz visiting // on da smartphone,” illustrates that poems in Hawaiian Islands Pidgin (HIP) can be challenging for machines to understand due to the author’s intentional misspellings. In contrast, the author’s note complements it: “The idea for this poem came from an actual phone call that I had with my sister," features correct spelling and a more straightforward style, allowing machines to understand the theme more easily. Likewise, despite its brevity, the about-poem data provides additional information in a more interpretable fashion, leading us to investigate its utility for the task in this paper.

The poems are labeled with 156 unique themes, primarily representing the subjects of the poems. However, some themes do not correspond to the poem's subject matter, such as public domain, audio, and translation, which we removed. Each poem can have multiple themes associated with it, with an average of 2 themes linked to a single poem. Hence, our classifiers are trained to predict an unspecified number of thematic classes per poem. Table 1 displays the top 50 most prevalent themes among the poems with the about-poem data, accompanied by their respective poem counts. These themes cover a broad spectrum of topics related to human experience, including both abstract (e.g., love, existential, etc.) and concrete subjects (e.g., animals, flowers, etc.).

Among the 156 themes, we employ two separate subsets based on their popularity: the top 10 and top 50 themes. This selection was made to ensure comparability with previous studies on poem topic classification by incorporating an equal number of themes (Lou et al., 2015).

Theme (Count)
nature (467), body (453), death (353), love (350), self (348), existential (334), identity (334), america (292), beauty (269), memories (264), ancestry (235), animals (220), time (208), loss (204), history (201), landscapes (193), thought (184), writing (182), family (174), violence (166), desire (155), grief (141), social justice (141), aging (139), flowers (136), language (133), childhood (132), past (127), birds (127), music (127), environment (126), mothers (123), night (119), politics (119), hope (114), oceans (111), cities (104), war (104), gender (103), fathers (102), earth (102), survival (100), religion (100), heartache (97), dreams (95), loneliness (94), illness (92), weather (90), spirituality (87), anxiety (86)

Table 1. Top 50 Popular Themes in Poems with About-Poem Data and Their Respective Counts.

Classification Setup

In this paper, we employ BERT and SVM for multilabel-multimodal poem classification, extending their previous application in single-label classification tasks (Choi, 2021, Choi et al., 2016). This key structural modification reflects the fact that real-world poems can contain multiple themes simultaneously. In addition, we investigate a novel combination of the bimodal input pair (poem and the author's note), which is anticipated to be better suited to handle theme classification than the unimodal case (i.e., poem-only).

BERT (Devin et al., 2018) is a pre-trained natural language processing model based on the Transformer architecture (Vaswani et al., 2017). Unlike context-independent models, such as Word2Vec (Mikolov et al, 2013) and fastText (Bojanowski et al., 2017), BERT captures contextual information by learning embeddings for words in relation to their surrounding context within a sentence. This feature enables BERT to better understand semantic relationships between words and provides improved performance in various NLP tasks. BERT models are pretrained on large datasets and can be used effectively even with limited domain-specific data.

In our approach, illustrated in Figure 1, we utilize an ensemble method that combines two BERT classifiers, one for the poem and the other for the author's about-poem note, respectively, by performing a weighted average of the two modality-specific results. The ensemble weight, where $0 \leq \alpha \leq 1$, is a trainable parameter optimized during the model training process. We start from the pretrained BERT model available in the ktrain package (Maiya, 2022). The pretrained BERT model performs inference on the "CLS" token-appended input sequence. Subsequently, a softmax layer follows to capture the CLS token's embedding, which encodes discriminative features. This process is performed separately for both input modalities, with their results combined using the fusion weight. Due to computational constraints and the relatively small dataset size, we freeze the BERT layers; instead, the softmax layers of both modalities and their fusion weights are updated during training. We employ a one-cycle policy for optimization (Smith, 2018) and follow a five-fold cross-validation process to validate the results. The implementation is based on the Keras deep learning framework (Ketkar & Ketkar, 2017).

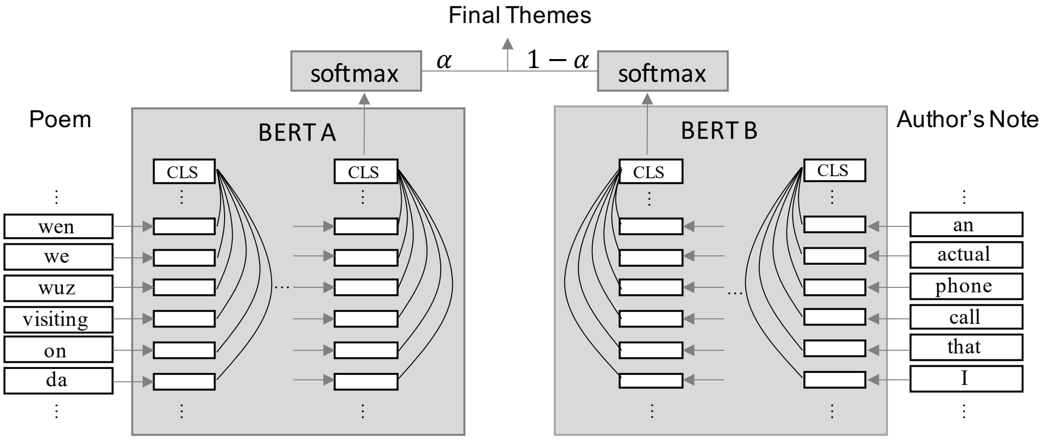

Figure 1. The Proposed Bimodal Classification System for Poems and Author's Notes: Integrating the Two Modalities with a Tunable Ensemble Weight α During Training

We also build an SVM-based baseline classification model with the ordinary TF-IDF representation, following the approach demonstrated by Lou et al. (2015). SVM has proven to be effective in a variety of studies on music and text classification (Hu et al., 2005; Yang & Chen, 2012). Based on Choi et al. (2016), which highlighted the robust performance of linear kernels in SVM models, we choose to use the linear kernel for the SVM model. We also employ the ensemble method, using two classifiers for the poem and the author's about-poem note, and combine their results with a trainable fusion weight, α, as in the BERT-based model. For implementation, we utilize the scikit-learn library (Pedregosa et al., 2011).

For the performance measure, we utilize the Area Under the ROC (Receiver Operating Characteristic) Curve, abbreviated as AUC, in accordance with previous multilabel classification studies (Lou et al., 2015; Choi et al., 2016). Unlike the conventional accuracy measure, which is primarily used in single-label classification tasks involving balanced datasets, AUC offers a more appropriate and robust single-value summary for multilabel classification in imbalanced dataset scenarios: the metric adjusts the number of theme classes per poem from the most conservative choice (i.e., a poem is classified into the most promising single class) to the most aggressive case (i.e., a poem belongs to all classes) to observe the sensitivity of the classifier. AUC values typically fall between 0.5 and 1.0, where a score of 0.5 signifies poor classifier performance and a score of 1.0 indicates perfect performance.

RESULTS AND DISCUSSION

This section presents a comprehensive analysis of the multilabel poem classification results, focusing on three main aspects: the comparison of BERT-based and SVM-based models, the comparison of three input types, and an analysis of the classification performance across various themes.

In comparing the BERT-based and SVM-based models, Table 2 demonstrates the superior performance of the BERT-based model across all input types and datasets. For both the 10-themes and 50-themes datasets, the BERT-based model achieved higher AUC scores than the SVM-based model in all cases, including poems, notes, and bimodal inputs. The low AUC score for the unimodal (poem-only) SVM model aligns with the findings of Lou et al. (2015), despite using a different poem dataset. The BERT model's superior performance may stem from its ability to leverage contextual information, which the SVM model struggles to capture with its TF-IDF input. Interestingly, the relatively small performance difference between the 10-themes and 50-themes datasets implies that the multilabel classification systems can effectively maintain performance even when addressing a larger number of themes.

System	The Proposed BERT-Based Model			The Baseline SVM		
Input	Poems	Notes	Bimodal	Poems	Notes	Bimodal
10-themes	0.77	0.74	0.80	0.72	0.68	0.75
50-themes	0.75	0.74	0.77	0.72	0.69	0.76

Table 2. Performance Comparison of BERT-Based and SVM-Based Models: AUC Scores for Different Input Types and Datasets

Among the three input types, bimodal models outperform unimodal models in both BERT- and SVM-based models, emphasizing the complementary nature of the poem and about-poem data. The estimated ensemble weight of the BERT-based model is not significantly different from a simple average: α = 0.50, with a nearly zero standard deviation for both datasets. This illustrates that both input data contribute equally to the performance improvement. Conversely, the ensemble weight of the SVM-based model reveals a greater contribution from poems, with α = 0.62 and a standard deviation of 0.07 for the 10-theme dataset and α = 0.84 with a standard deviation of 0.05 for the 50-theme dataset. Among the unimodal inputs, poem-only model outperforms the note-only model. The note-only model's lower performance, notwithstanding the benefit of its straightforward language, may be due to the brevity of the notes compared to the poems, as the notes constitute only 8.7% of the poems' length. Thus, we anticipate that incorporating additional auxiliary data, such as commentaries, could further improve performance.

Input	Nature	Body	Death	Love	Self	Exist.	Identity	Ameri.	Beauty	Mem.
Poem	0.86	0.80	0.76	0.86	0.73	0.68	0.76	0.83	0.72	0.73
Notes	0.79	0.75	0.70	0.82	0.66	0.68	0.77	0.84	0.70	0.67
Both	0.88	0.83	0.83	0.87	0.73	0.73	0.80	0.86	0.76	0.74

Table 3. Proposed Model’s Performance Across 10 Key Themes for Unimodal (Poems, Notes) and Bimodal (Both) Models, Using AUC Scores. Abbreviations: Exist. = Existential, Ameri. = America, Mem. = Memories.

Table 3 presents the proposed BERT model’s AUC values across 10 key themes for unimodal (poem and note) and bimodal (both) models. The bimodal model consistently demonstrates strong performance in classifying themes such as nature, love, America, body, and death. Conversely, it demonstrates weaker performance in themes such as existential, self, and memories, perhaps due to their more abstract nature and the potential overlap among these correlated themes. They share similar vocabulary and concepts, which makes accurate differentiation more challenging for the model.

CONCLUSION

In conclusion, this article addressed the scarcity of diverse metadata in digital poetry libraries and proposed a multilabel poem topic classification algorithm that leveraged large language models and auxiliary data. The study demonstrated the potential of context-dependent language models, such as BERT, for understanding poetic words. It also showed that incorporating auxiliary data, like author's notes, alongside these models can substantially improve the performance of the classification system, with an AUC increase from 0.72 to 0.80, despite the challenging nature of the task. The results suggested that multilabel classification systems maintain performance while handling a large number of themes. The study highlighted the importance of utilizing advanced technology, such as large pretrained language models, to enhance the accessibility of various poems within collections. With continued performance improvement, this approach will be valuable for discovering works from new authors or uncovering hidden gems with limited or no metadata, promoting a more inclusive representation of poetry. Ultimately, this research can assist in fostering the therapeutic benefits of poetry and enhancing the understanding and appreciation of the art form.

ACKNOWLEDGMENTS

This work was supported by RE-252382-OLS-22 from the Institute of Museum and Library Services.

REFERENCES

Academy of American Poets. (n.d.). About us. Poets.org. Retrieved April 16, 2023, from https://poets.org/academy-american-poets/about-us

Acim, R. (2021). Lockdown poetry, healing and the COVID-19 pandemic. *Journal of Poetry Therapy*, 34(2), 67-76.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, 135-146.

Choi, K. (2021). Bimodal Music Subject Classification via Context-Dependent Language Models. *In the 16th International Conference, iConference 2021*, Beijing, China, March 17–31, 2021, Proceedings, Part I 16 (pp. 68-77). Springer International Publishing.

Choi, K., Fazekas, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. *in Proceedings of Conference of The International Society for Music Information Retrieval*, 2016, pp. 805–811.

Choi, K., Lee, J. H., Hu, X., & Downie, J. S. (2016). Music subject classification based on lyrics and user interpretations. *Proceedings of the Association for Information Science and Technology*, 53(1), 1-10.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of NAACL-HLT* (pp. 4171-4186).

Harrower, M. (1972). The therapy of poetry.

Hu, X., Downie, J. S., West, K., & Ehmann, A. F. (2005, September). Mining Music Reviews: Promising Preliminary Results. *In ISMIR* (pp. 536-539).

Hynes, A. M. (2019). Bibliotherapy: The interactive process a handbook. *Routledge*.

Kao, J., & Jurafsky, D. (2012, June). A computational analysis of style, affect, and imagery in contemporary poetry. *In Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature* (pp. 8-17).

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).

Ketkar, N., & Ketkar, N. (2017). Introduction to keras. Deep learning with python: a hands-on introduction, 97-111.

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

Maiya, A. S. (2022). ktrain: A low-code library for augmented machine learning. *The Journal of Machine Learning Research*, 23(1), 7070-7075.

Mazza, N. (2021). Poetry therapy: Theory and practice. *Routledge*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Navarro-Colorado, B. (2018). On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry. *Frontiers in Digital Humanities*, 5, 15.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

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).

Richardson, L. (2007). Beautiful soup documentation.

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. *arXiv preprint* arXiv:1803.09820.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, Y. H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3), 1-30.