Poetry Theme Classification

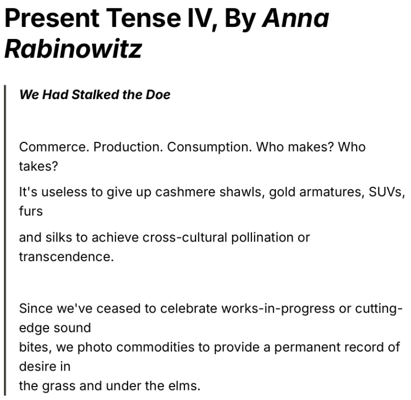
Pei-Chi Pan, Haonan Wang
Department of Computer Science, Rice University
pp37@rice.edu, hw82@rice.edu

Abstract

This project utilizes advanced natural language processing (NLP) techniques to classify poetry based on thematic elements, enhancing personalized exploration of poetic works. Utilizing a comprehensive dataset from Kaggle, this study employs a range of machine learning and deep learning models, including Logistic Regression, Random Forest, XGBoost, LSTM Neural Networks, and the BERT model, which is adapted specifically for poetry. Our approach includes preprocessing the textual data through tokenization and vectorization, and constructing pipelines for each model to evaluate their efficacy in thematic classification. The performance of these models is assessed using standard metrics such as accuracy, precision, recall, and F1-score, with deep learning models showing particular effectiveness in capturing the nuanced themes of poetry. The findings enhance the understanding of computational methods in literary analysis and highlight the potential of machine learning and deep learning in poetry classification.

1. Introduction

The Poetry Foundation dataset [1] is used for our model, presenting a challenge to accurately predict the thematic categories of each poem. Through meticulous data preprocessing efforts—ranging from handling missing values to normalizing text data—we prepare the ground for the application of sophisticated modeling techniques. The centerpiece of our approach is the utilization of the pre-trained BERT model [2], renowned for its deep bidirectional understanding of context, which offers a promising avenue for capturing the nuanced expressions of poetry.



Present Tense IV, By *Anna Rabinowitz*

We Had Stalked the Doe

Commerce. Production. Consumption. Who makes? Who
takes?
It's useless to give up cashmere shawls, gold armatures, SUVs,
furs
and silks to achieve cross-cultural pollination or
transcendence.

Since we've ceased to celebrate works-in-progress or cutting-
edge sound
bites, we photo commodities to provide a permanent record of
desire in
the grass and under the elms.

Figure 1. Example of a poem from the dataset. Tags: *Living, Growing Old, Social Commentaries, Money & Economics*.

2. Related Work

The integration of computational methods and literary analysis has significantly evolved, particularly with the application of natural language processing (NLP) to poetry. Historically, research has concentrated on stylistic analysis, authorship attribution, and emotional content extraction, demonstrating NLP's potential to uncover new insights into literary texts [4].

Stylistic Analysis and Authorship Attribution: This area explores the quantification and analysis of stylistic elements in poetry using machine learning to identify distinctive features of authors' styles and differentiate between literary periods.

Emotional Content and Imagery: Research in this domain utilizes sentiment analysis and image recognition to analyze the emotional and visual components of poems, aiding in thematic classification.

Despite these advancements, the thematic classification of poetry remains underexplored, particularly in addressing the abstract and subjective nature of poetic themes. The advent of advanced language models like BERT [3] introduces new capabilities for text analysis, necessitating adaptations for the complexities of poetic language, such as its symbolic

[1] https://www.kaggle.com/datasets/tgdivy/poetry-foundation-poems

nature and layered meanings.

Our project leverages these technological advancements, particularly focusing on the thematic classification of poetry using BERT, to enhance our understanding of thematic expressions in poetry and provide tools for deeper literary exploration.

3. Data Preprocessing

Our dataset, initially containing approximately 13,854 poems from a comprehensive Kaggle repository, was meticulously prepared through several steps. To ensure high data quality and usability, we implemented several cleaning measures

3.1. Data Cleaning and Normalization

First, we removed unnecessary whitespace and punctuation. Secondly, we handled missing data by excluding entries without 'Tags'. Lastly, we standardized all textual data by converting it to lowercase.

3.2. Focused Dataset Curation

Recognizing the importance of data relevance and representation, we refined our dataset by focusing on specific subsets of the data:

- **Selection of Prominent Poets:** We analyzed the distribution of works across poets and selected poems from the top 20 poets with the highest number of contributions. This focus ensured that our dataset included poems from prolific contributors, providing a rich basis for thematic analysis.

- **Tag Analysis and Simplification:** We examined the frequency of tags associated with the poems, selecting the top 20 most common tags for further analysis. This step was crucial for managing the complexity of the dataset and ensuring a balanced representation of themes.

- **Restructuring Tags:** To streamline the tagging system and reduce the complexity for the model, we grouped similar tags and eliminated infrequently used tags. This restructuring created a more generalized and manageable set of themes for the model to classify.

3.3. Data Exploration and Visualization

Our preprocessing included extensive exploration and visualization to better understand the dataset's characteristics:

- **Visualization of Poet and Tag Distribution:** We created visualizations to illustrate the distribution of poets' contributions and the prevalence of tags within the dataset. These visualizations helped identify dominant themes and influential poets, informing our approach to modeling.

- **Poem Length Analysis:** We computed and visualized the lengths of poems to capture the variability and common patterns in poem structures. This analysis provided insights into the typical composition of the poems, which could influence their thematic classification.

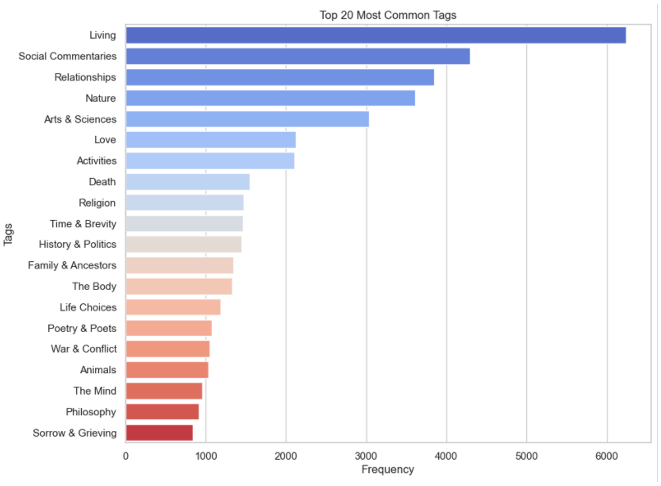


Figure 2. Top 20 tags distribution

4. Model

This study aims to develop an effective and efficient machine learning model for poem text classification. We built two types of classification systems, machine-learning based and deep learning-based system. For machine learning-based classification systems, we use TF-IDF and domain concept for feature extraction. The models here include XGBoost, Random Forest, and Logistic Regression. As for deep learning-based classification, we use LSTM Neural Network and Bidirectional Encoder Representations from Transformers (BERT).

The BERT model architecture combines the proven strengths of pre-trained models with specific adaptations for poetry theme classification. At its core, the encoder component utilizes a pre-trained BERT model, leveraging its deep bidirectional understanding of language context. BERT's pre-training on a vast corpus of text enables it to capture a wide range of linguistic nuances, making it an ideal foundation for understanding the complex and often abstract language found in poetry.

The decoder, in contrast, is not a traditional decoder as used in sequence-to-sequence models but rather a classification layer tailored to categorize the encoded poem into one of the thematic classes. This design choice reflects our task's nature, focusing on classification rather than generating textual output. The model's classification layer is fine-tuned on our dataset, allowing it to adapt the general lan-

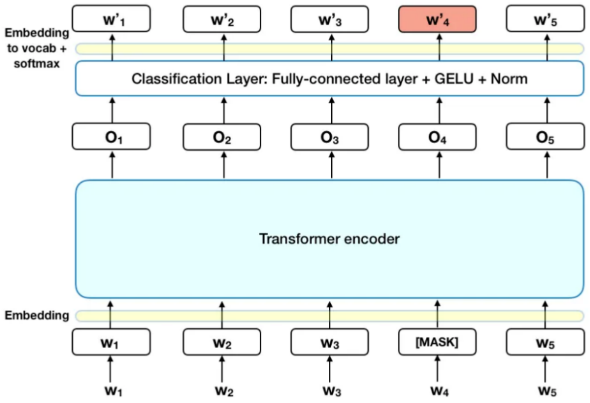

Embedding to vocab + softmax
w'1
w'2
w'3
w'4
w'5
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Figure 3. The BERT model is a multi-layer bidirectional Transformer en- coder. Its arquitecture is completely based in the model described in [1].

guage understanding of BERT to the specific thematic nuances of poetry.

5. Experiments and Results

5.1. Methodology and Experiment Design

In this study, we evaluated five different models to classify poetry based on thematic elements. The models tested were Logistic Regression, Random Forest, XGBoost, Neural Network, and BERT. Our evaluation focused on each model's ability to correctly predict the thematic categories of poetry, using a confusion matrix as the primary tool for analysis.

5.2. Model Performance

We have 20 categories in total but for evaluating the model performance, we only choose top 5 categories to evalute.

- **Logistic Regression:** This model showed a strong preference for certain themes such as ”Living,” but it struggled with ”Arts & Sciences” and ”Nature,” possibly due to their less frequent occurrence and higher thematic overlap with other categories.

- **Random Forest:** Performed well in distinguishing the most common themes but confused ”Nature” with ”Relationships,” suggesting that these categories might share linguistic similarities that the model could not differentiate.

  The general performance of each model can be found in Table 5.2.

- **Neural Network:** Offered a balanced performance across all categories, handling the nuances of thematic content with moderate success.

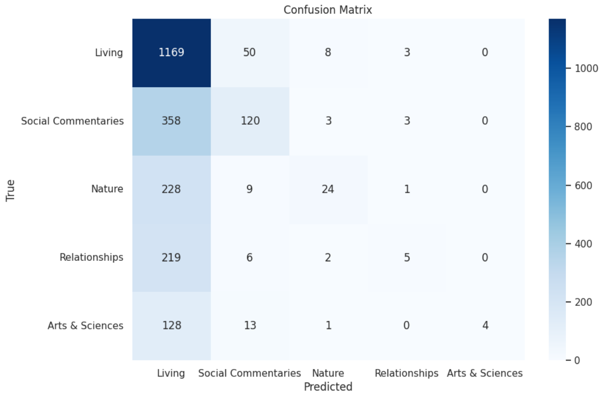


Figure 4. Logistic Regression Confusion Matrix.

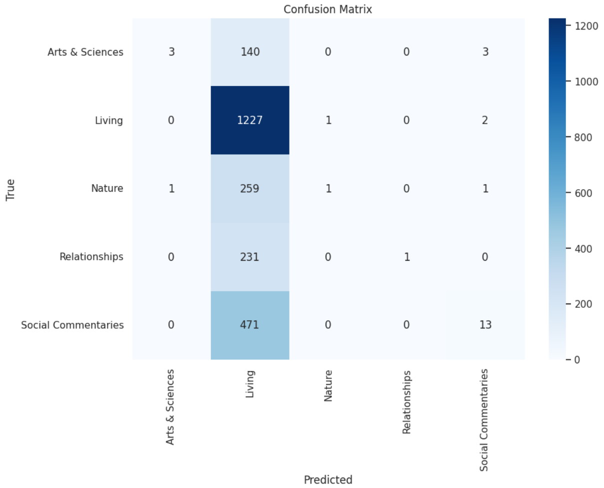


Figure 5. Random Forest Confusion Matrix. We can see blind predictions on the Living label due to data imbalance (%52 of ppoems correspond to the *Living* label).

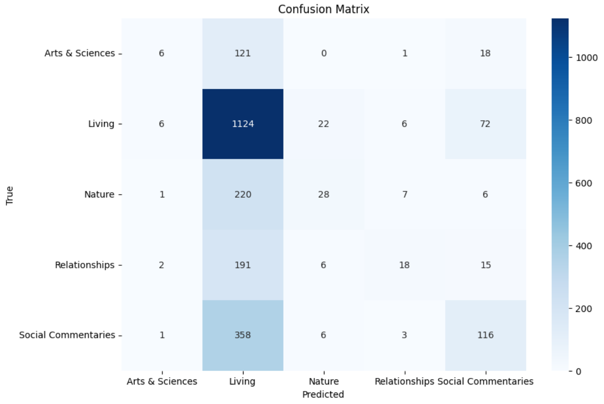


Figure 6. XGBoost Confusion Matrix. Marked improvement in handling less frequent categories compared to Logistic Regression and Random Forest, although still underperforming for ”Arts & Sciences.”

Model	Accuracy (%)	Precision (weighted)	Recall (weighted)	F1-score (weighted)
Logistic Regression	56.16	59%	56%	46%
Random Forest	52.89	62%	53%	38%
XGBoost	54.89	52%	55%	47%
Neural Network (LSTM)	41.00	40%	41%	40%
BERT	59.86	37%	39.47%	38.05%

Table 1. The table presents a comprehensive comparison of the performance metrics across different models employed in the study.

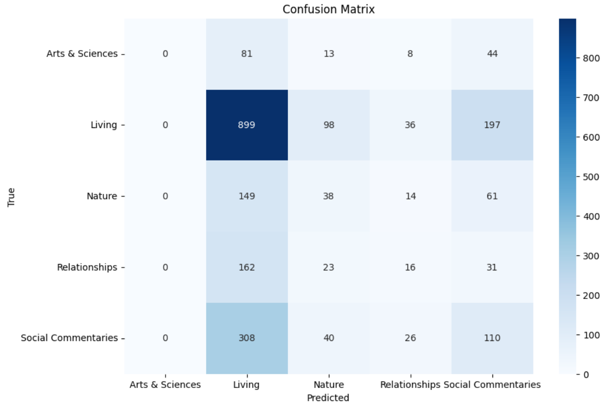


Figure 7. LSTM Neurual Network Confusion Matrix.

- **BERT:** Demonstrated superior performance, particularly in accurately classifying "Living" and "Social Commentaries," showcasing its capability to understand deeper linguistic contexts and subtleties.

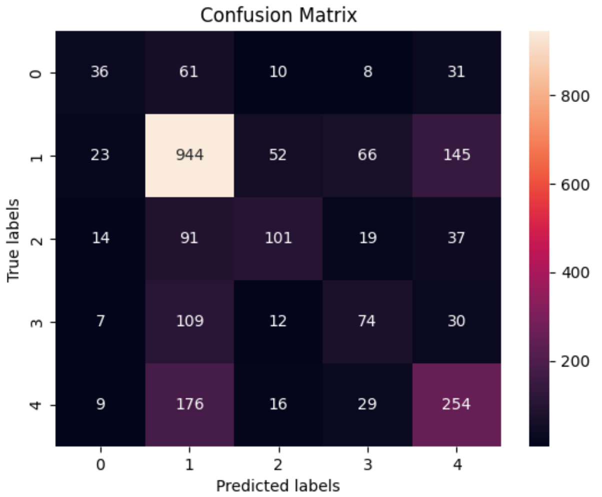


Figure 8. BERT Confusion Matrix. Label 0, 1, 2, 3, 4 represent Arts & Sciences, Living, Natural, Relationships, Social Commentaries respectively.

5.3. Discussion of Results

On the one hand, the results highlight the varying capabilities of each model in thematic poetry classification. BERT's leading performance underscores its advanced language processing abilities, making it particularly suitable for tasks involving complex language patterns and nuanced thematic distinctions. On the other hand, the prediction of living is more successful among all the models, due to the imbalance of the tag distribution. To further enhance the model, the issue is supposed to be fixed.

6. Conclusions

Employing advanced NLP techniques such as the pretrained BERT model is used to address the challenge of classifying literature by labels, in particular the Poetry Foundation dataset. This method results in an effective and accurate identification for this dataset. While BERT shows promising results, such as a higher accuracy in the model performance against other models (See Table 1.), it presents a lower performance in other measurements such as the precision and recall v.s. other models. These results highlight the complexity and nuanced nature of poetry for text classification.

The modest success of different models (e.g., Logistic Regression, Random Forest, LSTMs, etc.) underscores the inherent complexity in text classification based on interpreting the abstract and stylistically rich expressions found in poetry. This suggests that further advancements in model sensitivity, architecture, and training methodologies are needed to improve the thematic classification accuracy.

Moreover, future improvements should focus on a better fine-tuning of the model while incorporating vast linguistic features to enhance the understanding of poetic nuances, styles, metrics, and most importantly, usage of different formative techniques such as figurative, rhetorical and literal language. Additionally, balancing the dataset as well as implementing data balance and resembling, and integrating domain-specific knowledge could potentially improve model performance and generalization (See Fig. 4).

Finally, this research illuminates the challenges and potential of using NLP techniques for literary analysis and classification, setting a foundation for future works aimed at capturing the subtleties and vastness of styles in poetry.

7. Github Repository

https://github.com/Leon3331/COMP646-Project

References

[1] N. P. Ashish Vaswani, Noam Shazeer. Attention is all you need. 2017.

[2] J. Devlin, M.-W. Chang, and Lee. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] M.-W. C. Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2018.

[4] J. Kao and D. Jurafsky. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, pages 8–17, 2012.