Poetry Theme Classification

Pei-Chi Pan, Haonan Wang
Department of Computer Science, Rice University
pp37@rice.edu, hw82@rice.edu
https://github.com/Leon3331/COMP646-Project

Abstract

This project endeavors to classify poetry based on its thematic elements by leveraging advancements in natural language processing (NLP). With the aim of categorizing poems into distinct themes, we seek to facilitate personalized exploration and discovery of poetic works that resonate with individual interests. Utilizing a comprehensive dataset from Kaggle, containing poems tagged with respective themes, we harness the power of NLP techniques to unlock new pathways of exploration within the realm of poetry. Through meticulous data preprocessing and the application of sophisticated modeling techniques, we aim to illuminate the thematic richness of poetry and enhance accessibility for readers.

1. Introduction

The dataset serves as the foundation 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 [1], renowned for its deep bidirectional understanding of context, which offers a promising avenue for capturing the nuanced expressions of poetry.

2. Related Work

The intersection of computational methods and literary analysis has garnered increasing interest over the past few decades, with several pioneering studies setting the stage for the application of natural language processing (NLP) in the field. In the specific realm of poetry analysis, previous research has mainly focused on stylistic analysis, authorship attribution, and emotional content extraction. These studies have demonstrated the potential of NLP to provide new insights into literary texts.

Stylistic Analysis and Authorship Attribution: One of the foundational areas in computational poetry analysis is the study of stylistic features and authorship. Works such as those by Kao and Jurafsky [2] have explored how stylistic elements can be quantified and analyzed using machine learning techniques. These studies typically employ statistical methods to identify unique features of authors' styles or to distinguish between different literary periods.

Emotional Content and Imagery: Another significant area of research focuses on the emotional content and imagery within poems. Studies have used sentiment analysis and image recognition techniques to analyze how emotions and visual imagery are conveyed through text. This line of research has helped in understanding the affective layers of poems, which are crucial for thematic classification.

Despite the progress in these areas, the thematic classification of poetry, which is the focus of our project, remains a relatively underexplored field. Most existing studies have not fully addressed the challenges posed by the abstract and subjective nature of themes within poetry.

Advancements in Language Models: More recently, the advent of advanced language models like BERT [3] has opened up new possibilities for text analysis. These models are pre-trained on large text corpora and fine-tuned for specific tasks, allowing for a deeper understanding of context and nuance. However, their application to poetry requires adaptations to handle the unique challenges posed by poetic language, such as its condensed form, symbolic language, and layered meanings.

Our project builds upon this existing body of work by focusing specifically on the thematic classification of poetry using the BERT model. By adapting this model to the specific nuances of poetic text, we aim to advance the understanding of how themes are expressed and perceived in poetry. This effort not only contributes to the computational analysis of literature but also provides a practical tool for readers and educators to explore poetry through a thematic lens.

3. Data Preprocessing

Effective data preprocessing is crucial for the success of any machine learning project, particularly when dealing with the complex and varied data involved in poetry analysis. Our dataset, sourced from a comprehensive Kaggle repository, includes 14,000 poems, most tagged with multiple themes. This section outlines the detailed preprocessing steps we undertook to prepare this dataset for the thematic classification task.

3.1. Data Cleaning

The initial step in our preprocessing pipeline involved cleaning the data to ensure its quality and usability. This process included:

- **Whitespace Removal:** Stripping unnecessary whitespace from text fields, such as the 'Tags' and 'Poem' columns, to prevent any parsing errors during the model training.
- **Missing Data Handling:** Removing rows where essential fields like 'Tags' were missing. This step is crucial as missing tags would provide no target for training our model.
- **Standardization:** Normalizing text data by converting all text to lowercase and removing punctuation, which helps in reducing the complexity of the model's input space.

3.2. Tag Analysis

Given the multi-tagged nature of the dataset, our next task was to analyze and structure the tags appropriately:

- **Tag Distribution Analysis:** Examining the distribution of tags to identify any imbalances. We found that some themes were overrepresented, while others had very few associated poems.
- **Tag Simplification:** To manage the complexity, we simplified the tagging system by grouping similar tags together and eliminating rarely used tags, ensuring a more balanced dataset for training.

3.3. Sampling and Partitioning

To address the issues of tag imbalance and the multi-tagged nature of poems:

- **Stratified Sampling:** We extracted a stratified sample based on the frequency of themes, ensuring that each theme was adequately represented in our training set. This approach also helped to mitigate the impact of imbalanced data on our model's performance.
- **Single-Tag Assignment:** Each poem was assigned a single, primary tag for the purpose of training. This was done by selecting the most dominant or frequent tag from the multiple tags associated with each poem, simplifying the classification task.

These preprocessing steps were designed to create a clean, balanced, and representative dataset that could effectively support the training of our NLP model, addressing both the technical challenges and the unique characteristics of poetic text.

> *"We Had Stalked the Does*
> *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.*
>
> *It's useless to give up cashmere shawls, gold armatures, SUVs, furs..."* - The Death of Atahualpa, Anna Rabinowitz

Figure 1. Poem extracted from the Poetry Foundation dataset, consisting of $14,000$ poems classified of tags (i.e., Art, science, religion, folklore, mythology).

4. Model

Our model architecture innovatively 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 Bidirectional Encoder Representations from Transformers (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 language understanding of BERT to the specific thematic nuances of poetry.

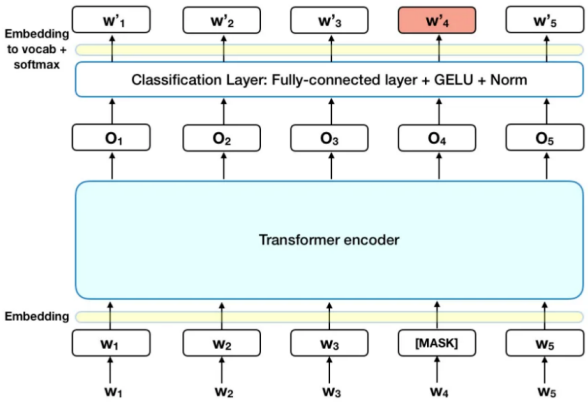

Figure 2. We use a pretrained model based on the **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) model. The difference between this model and other recent language representation models (Peters et al., 2018a; Rad- ford et al., 2018), is that BERT is designed to pre- train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a re- sult. BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

5. Experiments and Results

5.1. Training Process

- **Data Splitting**: The dataset was split into training (70%), validation (15%), and test (15%) sets. This partitioning ensured that we had separate data for training the models, tuning hyperparameters, and evaluating performance.

- **Optimization**: We used the ADAM optimizer for its adaptive learning rate capabilities, which is particularly beneficial for dealing with the varying scales of gradients in BERT.

5.2. Model Performance

- **Initial Findings**: Initially, the BERT model exhibited low accuracy, achieving only a 20% correct classification rate on the validation set. This was indicative of potential issues with overfitting to the training data or underfitting due to the complexity of the task.

- **Adjustments and Improvements**: To address these issues, we adjusted our approach by focusing the training exclusively on poems categorized under the top ten "Tags". This refinement in training scope helped improve model focus but yielded only modest gains in accuracy.

5.3. Result Analysis

- **Loss and Accuracy Trends**: Throughout the training process, we monitored the loss and accuracy metrics. Training and validation loss curves suggested slight overfitting, which is typical with high-capacity models like BERT when applied to nuanced tasks such as poetry classification.

- **Accuracy Improvement**: Even with adjustments, the model's accuracy did not exceed 20% on the validation set, underscoring the challenges inherent in classifying thematic content in poetry, which often involves abstract and subjective interpretations.

5.4. Discussion of Results

- **Insights and Implications**: The results highlight the difficulty of applying standard NLP techniques to literary texts, where thematic elements are deeply intertwined with artistic expression. These outcomes suggest a need for more specialized approaches, possibly incorporating broader contextual understanding or enhanced feature engineering tailored to literary analysis.

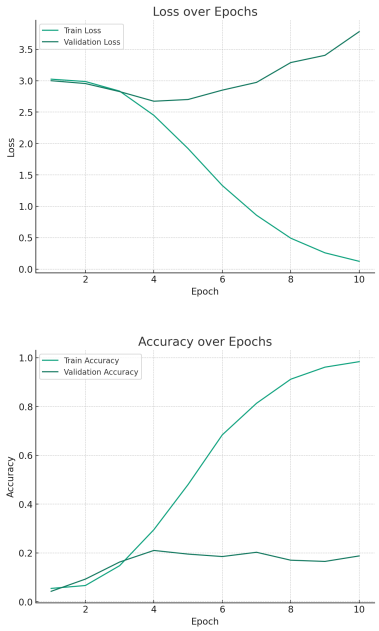

Figure 3. **Training Epochs vs. Loss/Accuracy**: Graphs illustrating the training and validation loss and accuracy provided visual insights into the model's learning behavior, indicating how well the model adapted to the complexity of the dataset over time.

6. Conclusions

In this project, we endeavored to apply advanced natural language processing techniques, specifically the pre-trained BERT model, to classify poetry based on thematic

elements. While our model achieved an accuracy of 20%, this result highlights the intrinsic challenges associated with the thematic classification of poetry—a domain marked by nuanced and abstract language.

This modest accuracy underscores the complexity of the task and indicates that poetry, with its rich and varied expressions, poses significant challenges that differ markedly from more conventional text classification tasks. Despite these challenges, our project provides valuable insights into the limitations and requirements of applying NLP techniques to literary texts.

7. Future works

To enhance the model's performance, we plan to explore data augmentation techniques and test different preprocessing strategies to address the complexity of poetic language. Experimenting with various model architectures may also offer improvements. Through these efforts, we aim to better adapt NLP technologies for literary analysis, ultimately making poetry more accessible and personalized for readers.

References

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

[2] 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.

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

A. Github Link

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