HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

*

FINAL PROJECT REPORT

Vietnamese poems category classification

Class ID: 152447

Lecturer: Đỗ Thị Ngọc Diệp

Group member: Tạ Lê Hiếu – 20224283

Nguyễn Chí Khiêm - 20224318

Hà Nội, tháng 1 năm 2024

TABLE OF CONTENTS

I. Overview

Poetry is a significant part of Vietnamese literature and culture, capable of expressing the author’s emotions while carrying layers of deep meaning. Categorizing poems can help organize literary works, enhance digital libraries for better accessibility, and provide readers with an easier way to grasp the essence of a poem. While this is an interesting task, it presents challenges due to the stylistic subtleties and cultural depth of poetry, along with the limited availability of annotated datasets.

This project aims to develop a model that classifies Vietnamese poems into predefined thematic categories using NLP techniques. The process involves text preprocessing, employing machine learning models, and evaluating their performance.

II. Investigating different methods

Method 1: Support Vector Machine (SVM) with TF-IDF

Description: SVM is a reliable classifier for handling high-dimensional data. It uses TF-IDF vectors to represent poems, highlighting the importance of words in individual poems compared to the whole dataset.

Advantages:

- Suitable for small to medium-sized datasets.
- Easy to interpret and implement.

Scientific Paper Reference: Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features.

Method 2: PhoBERT

Description: PhoBERT is a state-of-the-art pretrained language model specifically designed for Vietnamese text. It captures deep contextualized representations and provides superior performance in NLP tasks like text classification.

Advantages:

- Pretrained on a large corpus of Vietnamese text, ensuring contextual understanding.
- Outperforms traditional methods like LSTM and SVM on Vietnamese NLP tasks.
- Reduces the need for extensive feature engineering.

Scientific Paper Reference: Nguyen, D. Q., Nguyen, T. D., Pham, M. H., & Nguyen, A. T. (2020). PhoBERT: Pre-trained Language Models for Vietnamese.

Method 3: Multilayer Perceptron (MLP) with TF-IDF

Description: This method feeds TF-IDF vectors into an MLP, which is a fully connected neural network. The MLP learns relationships between the input features and target categories.

Advantages:

- Simple to design and train.
- Works well with carefully engineered features like TF-IDF.

Scientific Paper Reference: Multilayer perceptrons are a widely recognized architecture in neural networks, introduced in foundational works on artificial intelligence and machine learning.

Criteria	SVM with TF-IDF	PhoBERT	MLP with TF-IDF
Context Understanding	No	Yes	No
Performance on Small Data	Good	Good	Average
Resource Requirements	Low (sufficient with CPU)	High (better with GPU)	Medium (works with CPU/GPU)
Complexity	Low	High	Medium
Scalability	Medium	High	Medium

III. Methodology

1. Building and processing the Dataset

Our data source: http://vietnamthuquan.eu/Tho/

First and foremost, we developed a custom web scraping tool using Python and Selenium. This script automated the process of navigating through multiple pages of the website and extracting poems whose titles matched specific keywords. For further details, we recommend reviewing the second code cell in the crawling_process.ipynb file.

Key Features of the Tool:

a) Keyword Filtering:

- The tool filtered poem titles based on predefined keywords
- For **“Thơ thiên nhiên”**: ["xuân", "hạ", "thu", "đông", "núi", "biển", "hoa", "lá"]
  “Thơ đất nước”: ["quê", "quê hương", "đất nước", "tổ quốc", "nước", "đất"]
  “Thơ gia đình”: ["mẹ", "cha", "chị", "vợ", "chồng"]
  “Thơ tình”: ["anh", "em", "tình yêu"]
- This targeted approach helped streamline the dataset creation process by focusing on relevant content.

b) Automated Navigation:

- The tool utilized Selenium's WebDriver to open web pages, locate poem links, and navigate through multiple pages of the website.
- Dynamic page loading and "Next Page" navigation ensured all pages were processed efficiently.

c) Content Extraction:

- The script extracted poem titles and their content by locating specific HTML elements on each page.
- It employed multiple XPath queries to account for variations in the website's structure, ensuring comprehensive data collection.

d) Data Storage:

- Extracted data, including poem titles and content, was stored in a CSV file (poems_final.csv).
- This format was chosen for its compatibility with data preprocessing and machine learning pipelines.

e) Error Handling:

- Robust error handling mechanisms were included to manage issues such as missing elements or pages without poem content.
- Poems with incomplete data were flagged and logged for manual review.

Example Workflow:

- The tool started at a specific page of the website and iteratively navigated through the subsequent pages until no more pages were available.
- For each poem:
  - The title was checked against the keyword list.
  - If a match was found, the poem's content was scraped and stored in the CSV file.
- The script continued until all pages and matching poems were processed.

One major challenge we encountered was the overlap of keywords across multiple categories (ex. “Tình thu”). Additionally, some poems had titles that did not align with their content. As a result, every 20 minutes, the scraping process yielded around 100-150 poems, which required manual review to filter out irrelevant or mismatched data. However, as Vietnamese poems often carry multiple layers of meaning, it remains difficult to fully categorize a poem into a single thematic group.

After collecting sufficient data for each category-specific CSV file, we combined them into a single dataset and balanced it across all categories. The final dataset is well-prepared for subsequent preprocessing and model training steps.

Dataset Statistics:

- **Total Poems:** 2,400
- **Categories:**
    - *Thơ gia đình*: 600
    - *Thơ thiên nhiên*: 600
    - *Thơ tình*: 600
    - *Thơ đất nước*: 600

2. Implementation of the chosen Method

2.1. Selected Method: SVM with TF-IDF.

2.2. Dataset Preparation and Splitting

To ensure robust model training and evaluation, the dataset was carefully processed and split into training, testing, and validation subsets. The following steps were carried out:

a) Label Encoding:

Poem categories were converted into numerical labels using a label encoding technique. This process allowed the categorical target variable to be used effectively in machine learning algorithms.

b) Dataset Splitting:

The dataset was split into three subsets:

- **Training Set:** 80% of the data was reserved for training the model.
- **Testing Set:** 10% of the data was used to evaluate the model's generalization to unseen data.
- **Validation Set:** 10% of the data was allocated for hyperparameter tuning and model validation.

The splitting was performed using stratified sampling to ensure that each subset maintained the original category distribution.

c) Saved Datasets:

The resulting subsets were saved as separate CSV files for easy access and reproducibility:

- train_data.csv for training data.
- test_data.csv for testing data.
- val_data.csv for validation data.

Dataset Statistics:

- **Training Samples:** 80% of the dataset.
- **Testing Samples:** 10% of the dataset.
- **Validation Samples:** 10% of the dataset.

This structured approach to dataset preparation ensured a balanced distribution of categories across the subsets, facilitating reliable training and evaluation of the classification models.

2.3. Feature Extraction with TF-IDF

a) Input Data:

- X_train and X_test contain the poem content as raw text.
- Each poem is treated as a document.

b) Tokenization:

- The TfidfVectorizer splits each poem into individual words (tokens) by default (e.g., "xuân đến" → ["xuân", "đến"]).

c) Term Frequency (TF):

- Measures how often each word appears in a poem.
- Formula:

$$TF(t) = \frac{Number\ of\ occurrences\ of\ term\ "t"\ in\ a\ poem}{Total\ terms\ in\ the\ poem}$$

d) Inverse Document Frequency (IDF):

- Reduces the weight of words that appear frequently across all poems, as these are less likely to distinguish categories (e.g., "và", "là").
- Formula:

$$IDF(t) = \log\left(\frac{Total\ number\ of\ poems}{Number\ of\ poems\ containing\ term\ "t"}\right)$$

e) TF-IDF Calculation:

- Combines the two metrics:

$$TF - IDF(t) = TF(t) \times IDF(t)$$

- Words that are important in a specific poem but not common across all poems get higher weights.

f) Feature Matrix:

- **The TfidfVectorizer creates a sparse matrix where:**
  - Rows represent poems.
  - Columns represent unique words (features).
  - Values are the TF-IDF scores.

g) Feature Limitation:

- The max_features was set to 2500, meaning only the 2,500 most significant words (based on TF-IDF scores) are used as features. This reduces dimensionality and helps the model focus on important terms.

2.4. Training the SVM Model

We employed a **Support Vector Machine (SVM)** with a linear kernel for the classification task. The model was trained using the TF-IDF feature vectors generated from the training dataset. Key parameters include:

- **Kernel**: Linear, to efficiently separate poems into distinct thematic categories.
- **Regularization Parameter (C)**: Set to 1, balancing margin maximization and error minimization.

Once trained, the SVM model demonstrated robust capabilities in identifying and distinguishing between thematic categories of poems. Its performance was further evaluated on the test dataset to ensure accuracy and generalization.

3. Evaluation of the Model

Evaluation Metrics:

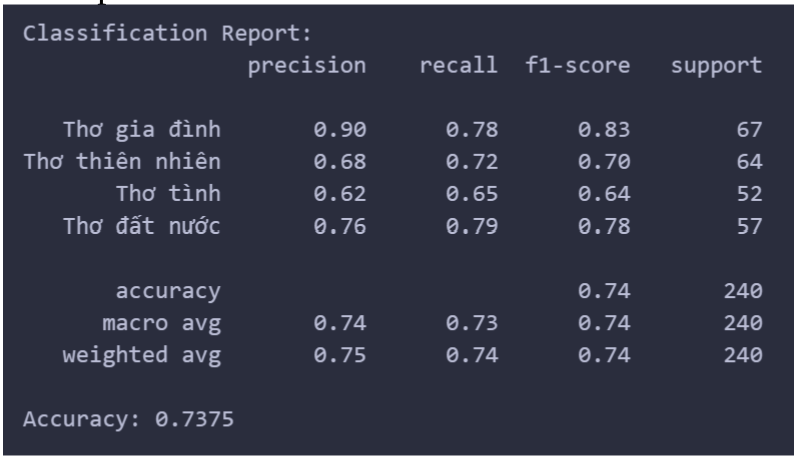
- **Accuracy:** Measures the proportion of correctly predicted categories.
- **Precision, Recall, and F1-Score:** Provide insights into the model's performance for each category.

Testing Process:

- The model was tested with unseen poems, both individually and in groups.
- **Example prediction:** A poem about family was accurately classified into the "Thơ gia đình" category.

Results:

The classification report and confusion matrix are illustrated below:



```
Classification Report:
                 precision    recall  f1-score   support

   Thơ gia đình       0.90      0.78      0.83        67
Thơ thiên nhiên       0.68      0.72      0.70        64
       Thơ tình       0.62      0.65      0.64        52
   Thơ đất nước       0.76      0.79      0.78        57

       accuracy                           0.74       240
      macro avg       0.74      0.73      0.74       240
   weighted avg       0.75      0.74      0.74       240

Accuracy: 0.7375
```

Fig1. Report of the SVM model

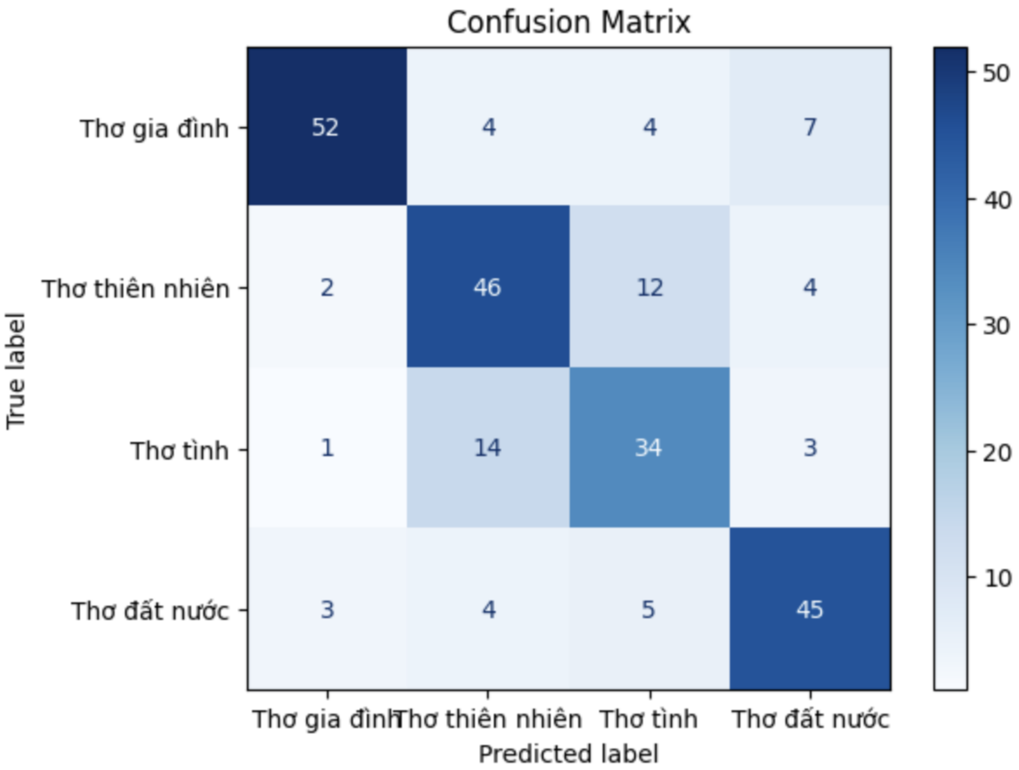


Fig2. The confusion matrix of SVM model

Analysis:

- The model performed well in distinguishing *Thơ gia đình* and *Thơ đất nước,* achieving high precision and recall scores.
- Categories like *Thơ thiên nhiên* and *Thơ tình* had overlapping vocabulary, leading to confusion between them.
- The overall accuracy was 73.75%, with a balanced F1-score across categories, indicating reasonable performance for a small dataset.

Future Improvements:

- Increase dataset size to improve the model's ability to distinguish between overlapping categories.
- Experiment with more advanced embeddings like PhoBERT to improve context understanding.
- Apply data augmentation techniques to balance the dataset across categories.

IV. Conclusion

This report examined the task of classifying Vietnamese poems into thematic categories using natural language processing (NLP) techniques. The Support Vector Machine (SVM) model with TF-IDF proved to be the most effective method, achieving an accuracy of 73.7%. It demonstrated a good balance between performance and computational efficiency, making it suitable for this project.

The study highlighted the challenges of working with poetry, such as overlapping themes and the complex, layered meanings inherent in Vietnamese literary works. Despite these difficulties, the SVM model performed well in distinguishing between thematic categories like Thơ gia đình and Thơ đất nước, while facing slight confusion in overlapping categories such as Thơ tình and Thơ thiên nhiên.

Overall, this project showcases the potential of applying simple yet effective machine learning techniques to thematic classification tasks in poetry. Future work could explore improving model performance by expanding the dataset, employing data augmentation techniques, or experimenting with advanced language models like PhoBERT. This research contributes to the broader goal of digitizing and organizing Vietnamese literary heritage for better accessibility and preservation.

--- THE END ---