Topic Modeling and Classification of Jin Yong Novels

Sanfeng Xin
by2410222@buaa.edu.cn

Abstract

Topic modeling is a crucial method for analyzing large textual corpora, with Latent Dirichlet Allocation (LDA) being widely used in literary studies. In this work, we apply LDA to the novels of Jin Yong to investigate how different preprocessing choices affect classification performance. Specifically, we analyze the influence of paragraph length, the number of topics, and the choice of text unit (words vs. Chinese characters) on the resulting topic distributions. Our experiments reveal that these factors significantly impact topic coherence and the ability to distinguish novels based on thematic structures. The findings provide insights into optimizing LDA parameters for classical Chinese literature and improving automated text classification in literary analysis.

Introduction

Latent Dirichlet Allocation (LDA)[1] is a widely used topic modeling technique that identifies latent themes in textual data. It has been successfully applied in various fields, including literature, social sciences, and biomedical text mining. However, the effectiveness of LDA depends on multiple factors, such as text segmentation, the number of topics, and the granularity of input text units. In Chinese natural language processing (NLP), a key choice is whether to use words or individual Chinese characters as the basic unit of analysis. Additionally, text segmentation at different levels, such as paragraphs or full documents, can affect the coherence and interpretability of LDA results.

In this study, we investigate the impact of paragraph length, topic granularity, and text unit selection (words vs. Chinese characters) on LDA-based classification of Jin Yong novels. By systematically varying these parameters, we evaluate how they influence topic coherence and the ability to distinguish different works. Our findings provide valuable insights into optimizing LDA for classical Chinese literature and contribute to computational literary analysis methodologies.

Methodology

Data Preparation

The dataset for this study consists of Jin Yong's sixteen martial arts novels, which serve as the textual corpus for topic modeling and classification. To ensure consistency and reliability in text processing, the following preprocessing steps were applied:

1.Remove useless content: The text was cleaned to remove any irrelevant content, such as advertisements at beginning and the end of the texts and blank characters. This step ensures that the analysis focuses solely on the narrative content of the novels.

2.Some novels have line break errors, and we have performed normalization.

3.Conversion from Traditional to Simplified Characters: The OpenCC [4] library was utilized to convert all traditional Chinese characters into their simplified counterparts. This step standardizes the script across the corpus, mitigating variations arising from orthographic differences (e.g., treating ”國” and ”国” as equivalent), which is crucial for maintaining consistency in character-level and word-level analyses.

4.Punctuation marks, special symbols, and numerical values unrelated to the content were removed.

5.Character-level analysis: Treating each Chinese character as an independent unit.

Word-level analysis: Using a Chinese word segmentation tool named jieba [2] to tokenize the text into meaningful words.

5.Removal of Stop Words: Frequently occurring stop words (e.g., ”的,” ”是,” ”在”) were removed from the tokenized text.

LDA Modeling

LDA is a generative probabilistic model that assumes each document is a mixture of topics, and each topic is characterized by a distribution over words. The model infers the latent topic structure from the observed data, allowing for the identification of thematic patterns within the corpus. In this study, we employed the Gensim library [3] to implement LDA modeling. The key parameters for LDA include:

Number of Topics and Length of Paragraphs

we vary the number of topics from 10 to 1000 and length of paragraphs from 20 tokens to 3000 tokens to build different LDA model,and calculate coherence scores for each model. Coherence score evaluates the interpretability of topics, with higher values indicating more coherent and meaningful topics. We use the UMass coherence measure [1] to assess topic coherence. The coherence score is calculated using the Gensim library, which provides an efficient implementation of the UMass coherence measure.

Words vs. Characters

In addition to the paragraph length and number of topics, we also investigate the impact of using words versus characters as the basic unit of analysis. This choice significantly influences the granularity of topic modeling and can affect the coherence and interpretability of the resulting topics. By comparing the results obtained from both word-level and character-level analyses, we aim to understand how this choice impacts the classification performance and topic coherence.

Comparison of Classifiers

To evaluate the effectiveness of different classification models in distinguishing Jin Yong’s novels based on LDA-derived topic distributions, we conducted a comparative study using multiple machine learning classifiers.

For this experiment, we fixed the LDA parameters as follows:

Number of Topics: 200

Paragraph Length: 1000 tokens

Feature Representation: Each document was represented as a 200-dimensional topic distribution vector (i.e., the probability distribution over topics).

We then applied six commonly used classification models to assess their performance:
Support Vector Machine (SVM)
k-Nearest Neighbors (KNN)
Multinomial Naïve Bayes (MultinomialNB)
Logistic Regression
Random Forest
Decision Tree

Conclusions

Impact of Topic numbers and Paragraph length

As paragraph length increases, the overall coherence of the LDA model tends to decline. This suggests that shorter paragraphs may better capture localized thematic structures, while longer paragraphs introduce more diverse content, making topics less distinct.

When using words as the unit of analysis, coherence is highest at a paragraph length of 100 tokens when the number of topics is 500 or 1000. This indicates that finer-grained segmentation, combined with a high topic count, helps maintain topic interpretability.

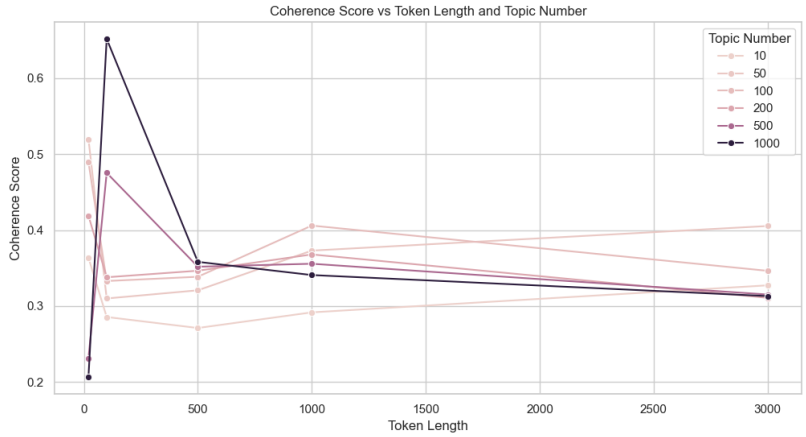

Figure 1: Coherence scores for different paragraph lengths and topic numbers using words as the unit of analysis.

When using Chinese characters as the unit, coherence varies based on the topic number:

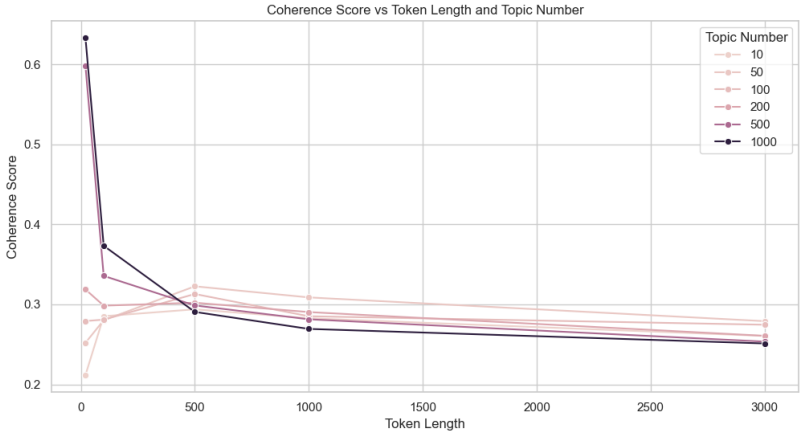

Figure 2: Coherence scores for different paragraph lengths and topic numbers using characters as the unit of analysis.

For fewer than 500 topics, coherence is highest at paragraph length of 500 tokens.

For 500 or more topics, coherence continues to decline as paragraph length increases, suggesting that longer paragraphs introduce too much variation for stable topic extraction.

Unlike topic coherence, classification accuracy improves as paragraph length increases, regardless of whether words or characters are used as the text unit.

Word-based segmentation: Increasing the number of topics improves classification accuracy, suggesting that a finer topic granularity provides better features for distinguishing novels.

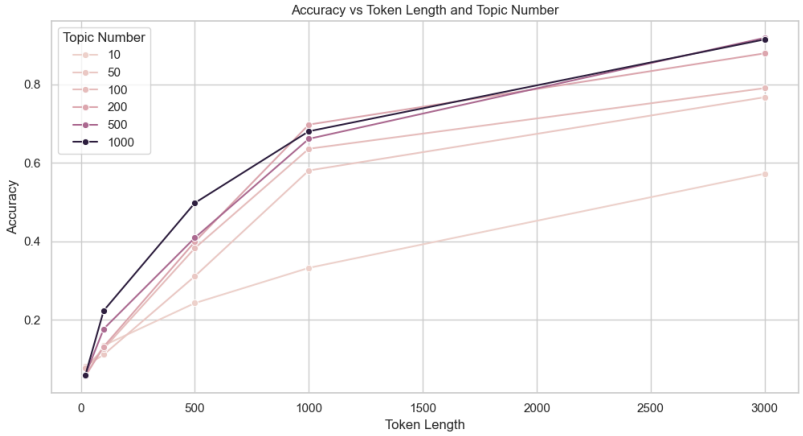

Figure 3: Classification accuracy for different paragraph lengths and topic numbers using words as the unit of analysis.

Character-based segmentation: The best classification accuracy is achieved with 200 topics, indicating that excessive topic granularity does not necessarily enhance classification performance when using characters as units.

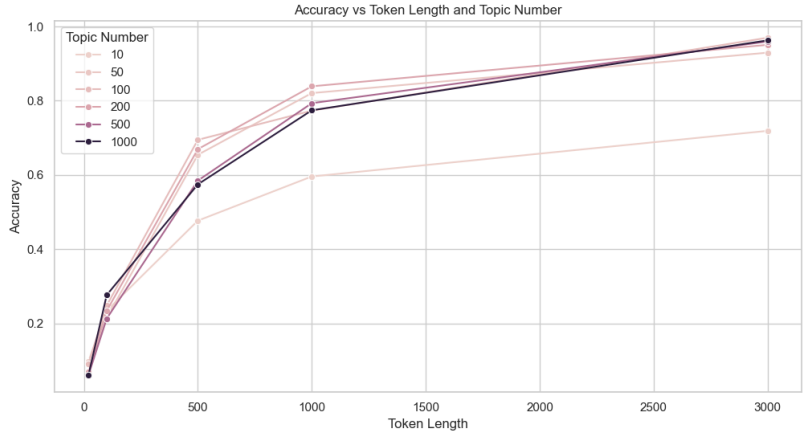

Figure 4: Classification accuracy for different paragraph lengths and topic numbers using characters as the unit of analysis.

Impact of Classifiers

We evaluated six classifiers (SVM, KNN, Naïve Bayes, Logistic Regression, Random Forest, and Decision Tree) on their ability to classify novels using LDA-generated topic distributions. The results show significant performance differences:

KNN performed the worst, achieving an accuracy of 0.55, indicating its limited effectiveness in handling topic-based feature representations.

Random Forest achieved the best performance, with an accuracy of 0.73, suggesting that ensemble methods are well-suited for LDA-based classification.

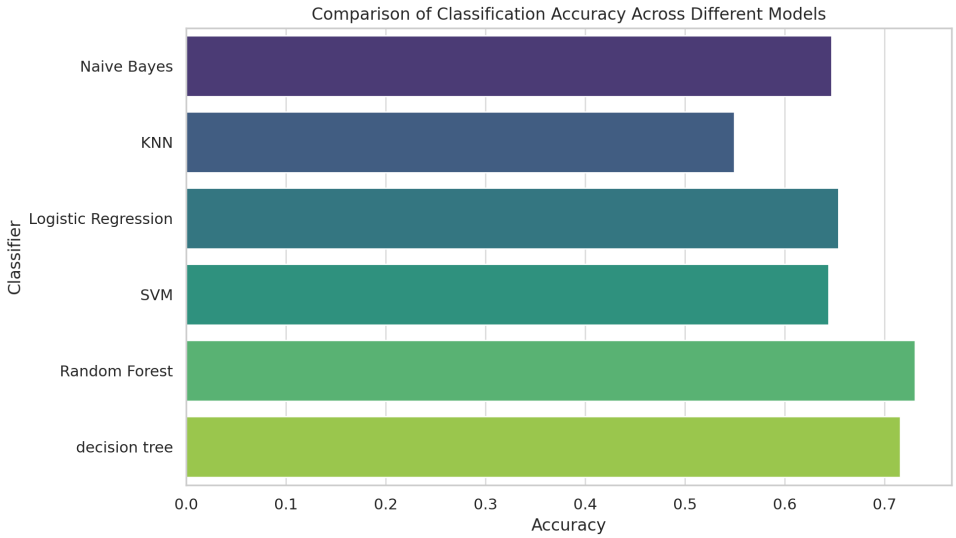

Figure 5: Classification accuracy of different classifiers using LDA-generated topic distributions.

References

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003.

[2] fxsjy. Jieba chinese text segmentation. `https://github.com/fxsjy/jieba`. Accessed: 2025-03-10.

[3] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[4] Xue Yang, Xiaowei Zhao, Gwan Tjio, et al. Opencc–an open benchmark data set for corpus callosum segmentation and evaluation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3020–3024. IEEE, 2020.