Comparative Analysis of Classical Polish Literature Using Small Language Models

Patrycja Wysocka, Tomasz Krupiński, Łukasz Jaremek, Mieszko Mirgos

Warsaw University of Technology

22.01.2025

Agenda

1. Introduction
2. Dataset generation
3. Methods
4. EDA
5. Experiments
6. Results and Analysis
7. Conclusions

Introduction

Research Overview

- **Problem:** Gap in understanding small LLMs' effectiveness for non-English literature
- **Approach:** Evaluation of 4 models on Polish classical texts
- **Innovation:**
  - Focus on smaller, accessible models (7B parameters)
  - Comprehensive evaluation framework

Dataset generation

Dataset Construction - Source Material

Corpus Composition:

- 11 seminal Polish works
- Mix of genres:
  - Poetry (Pan Tadeusz, Sonety)
  - Drama (Dziady, Kordian)
  - Novels (Lalka, Quo Vadis)
- Public domain sources (Wikisource)

Evaluation Task

- **Q&A Generation**
  - Question-Answer Pairs was a main task for all models.
  - Questions were generated using GPT-4
  - Duplicate questions were filtered out to ensure variety.
  - Books where splitted for 100 chunks and one question was asked to every chunk.
  - The final output consisted of Polish language question-and-answer pairs.
  - Generated questions focused on themes, characters, and key events in the texts.

Methods

Models Under Evaluation

- **General Multilingual:**
  - Qwen2.5 (7B parameters)
    - 128k context window
    - Broad domain coverage
  - LLaMA3.1 (8B parameters)
    - 128k context window
    - Instruction-tuned
  - Mistral v0.3 (7B parameters)
    - Designed for instruction-following tasks
    - High performance on multilingual benchmarks
- **Polish-Specific:**
  - Bielik (7B parameters)
    - 36B tokens training (22B Polish)
    - Mistral 7B base

Evaluation Framework

Primary Metrics:

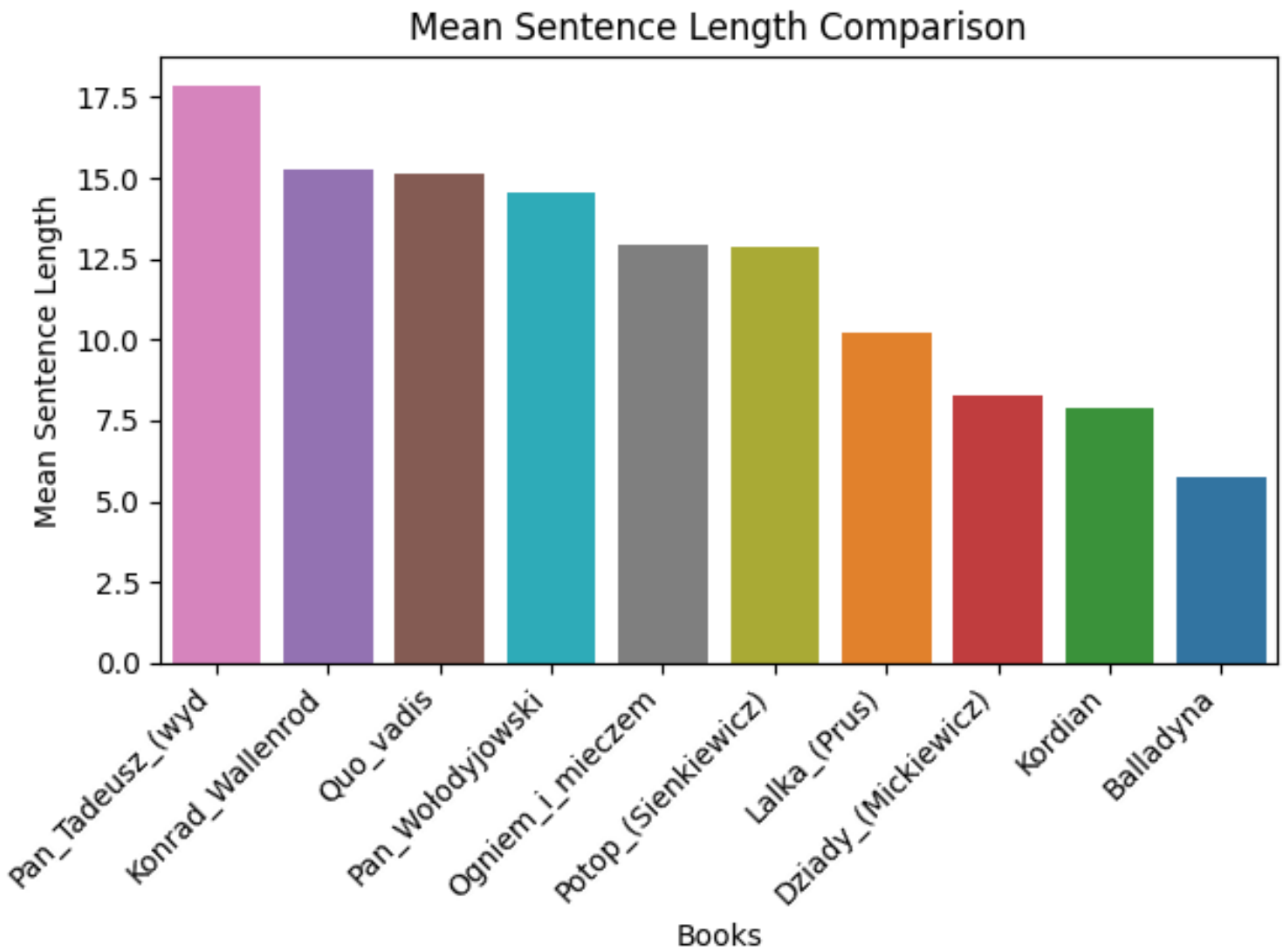
- **BLEU Score**
  - N-gram precision
  - Brevity penalty
- **METEOR**
  - Synonym matching
  - Word order evaluation
- **ROUGE Variants**
  - ROUGE-1

Followup Metric:

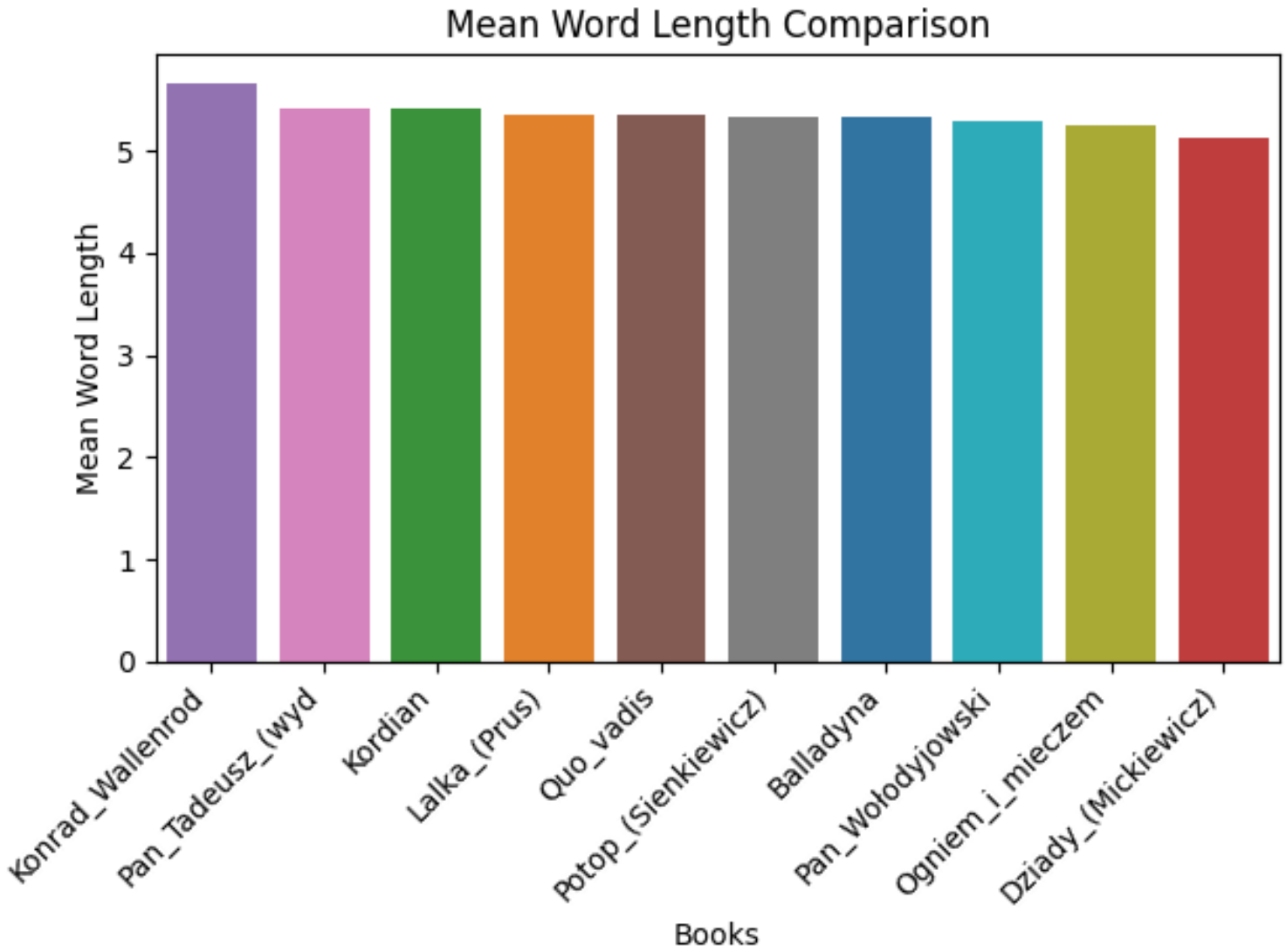
- **BERT-score**
  - Similarity evaluation
  - Transformer based evaluation

EDA

Mean Sentence Length Across Works

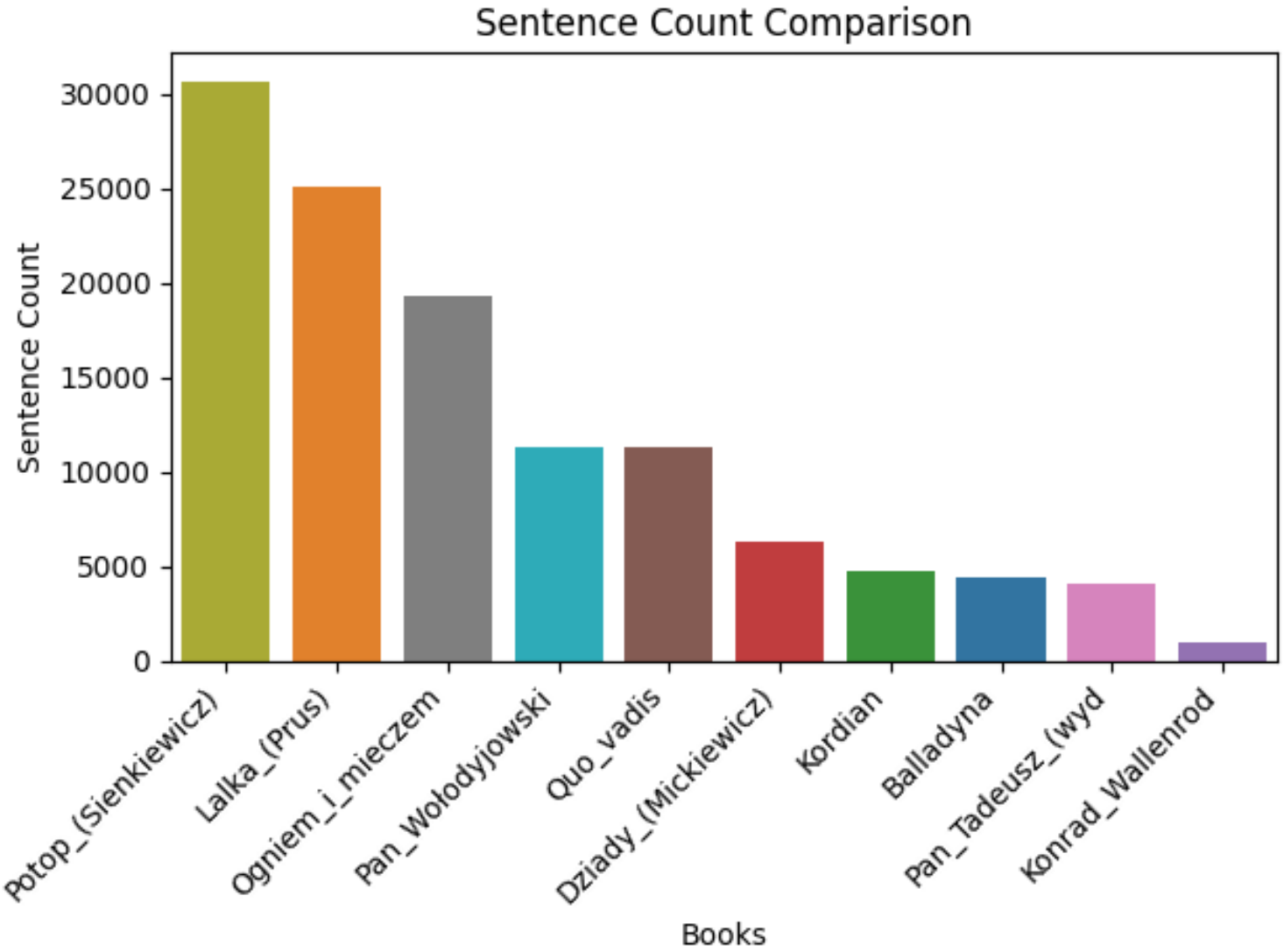


Mean Word Length



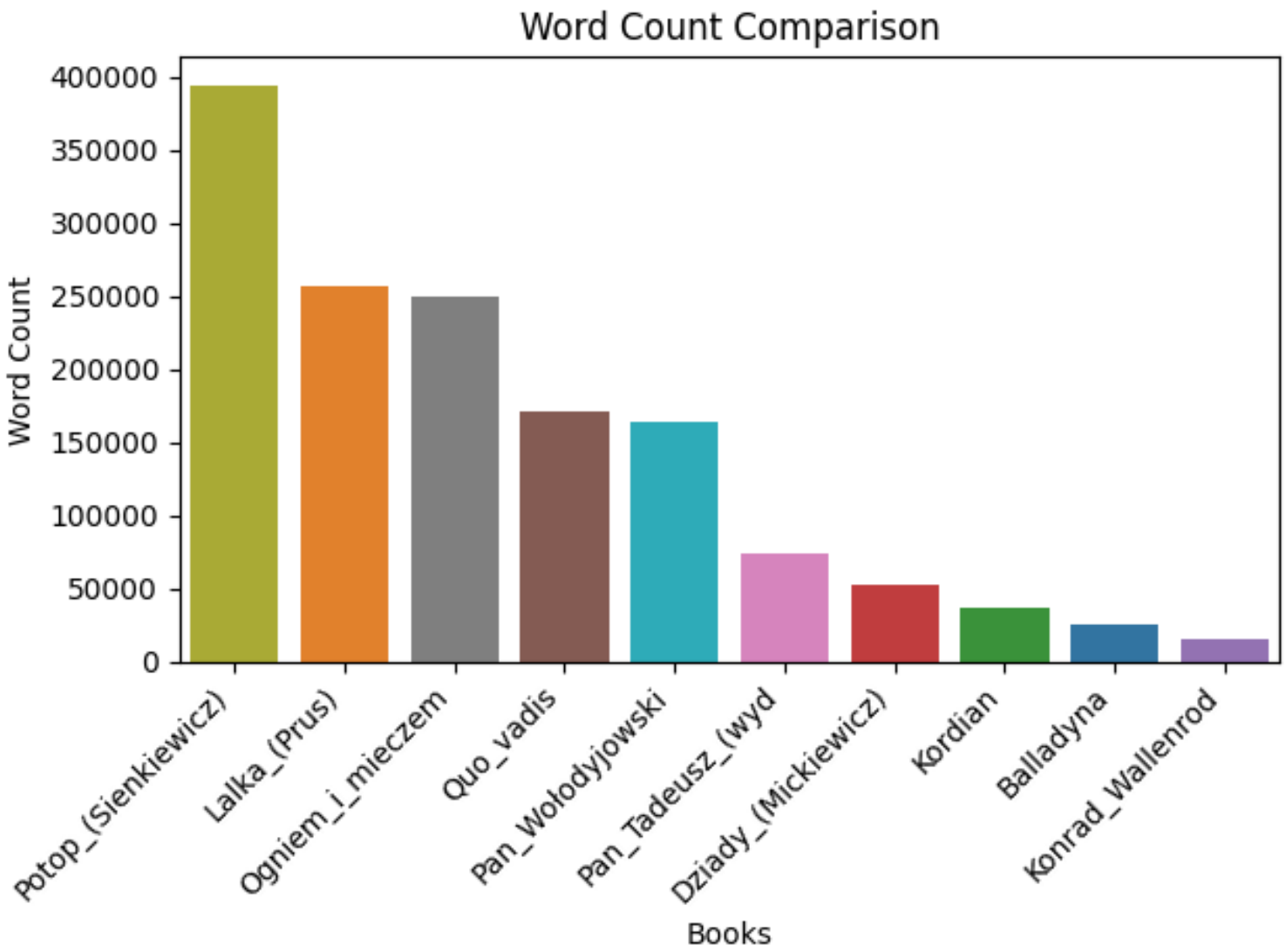
- Consistent across genres

Sentence Count Distribution



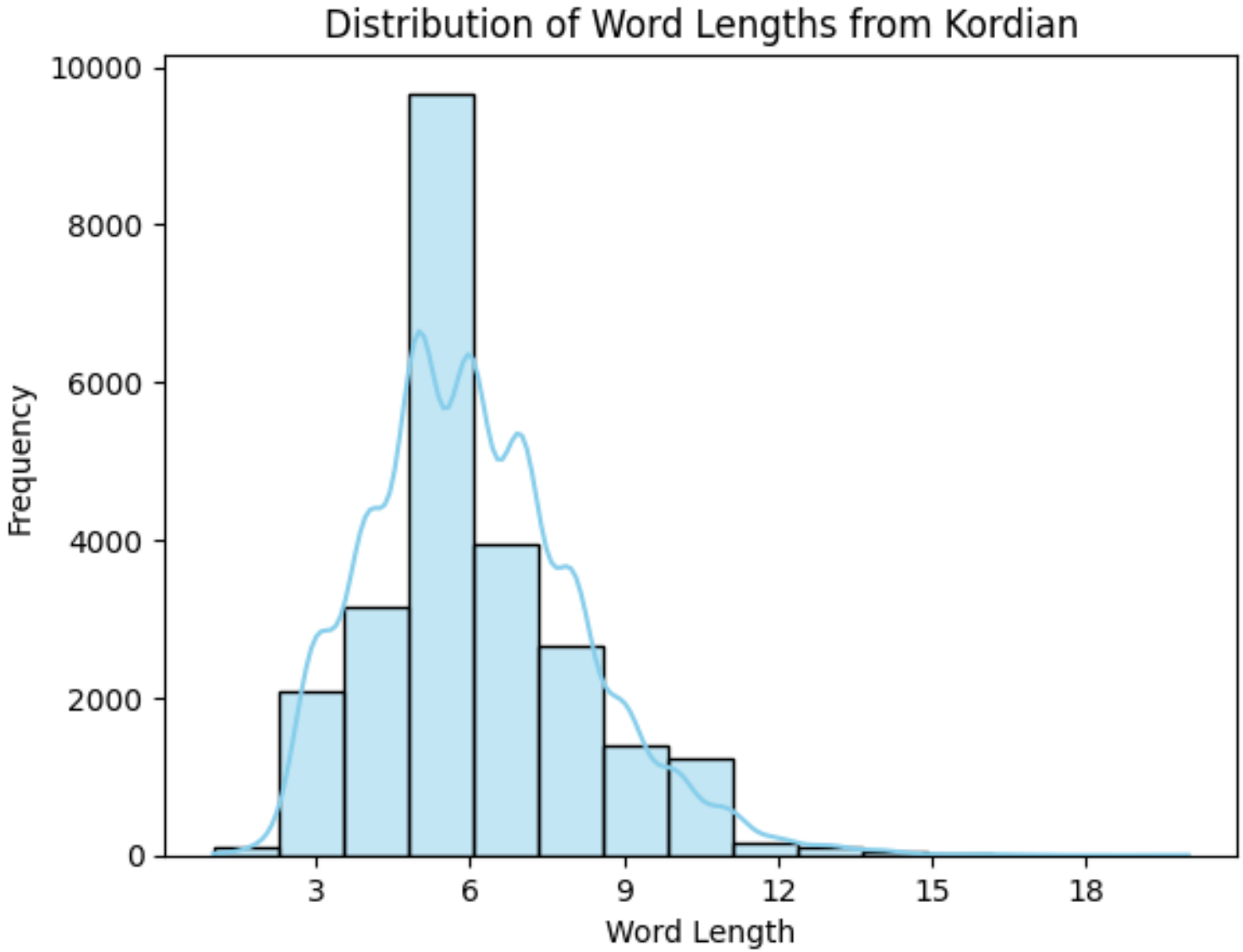
- Novels show highest sentence counts
- Poetry and drama more concise

Word Count Distribution



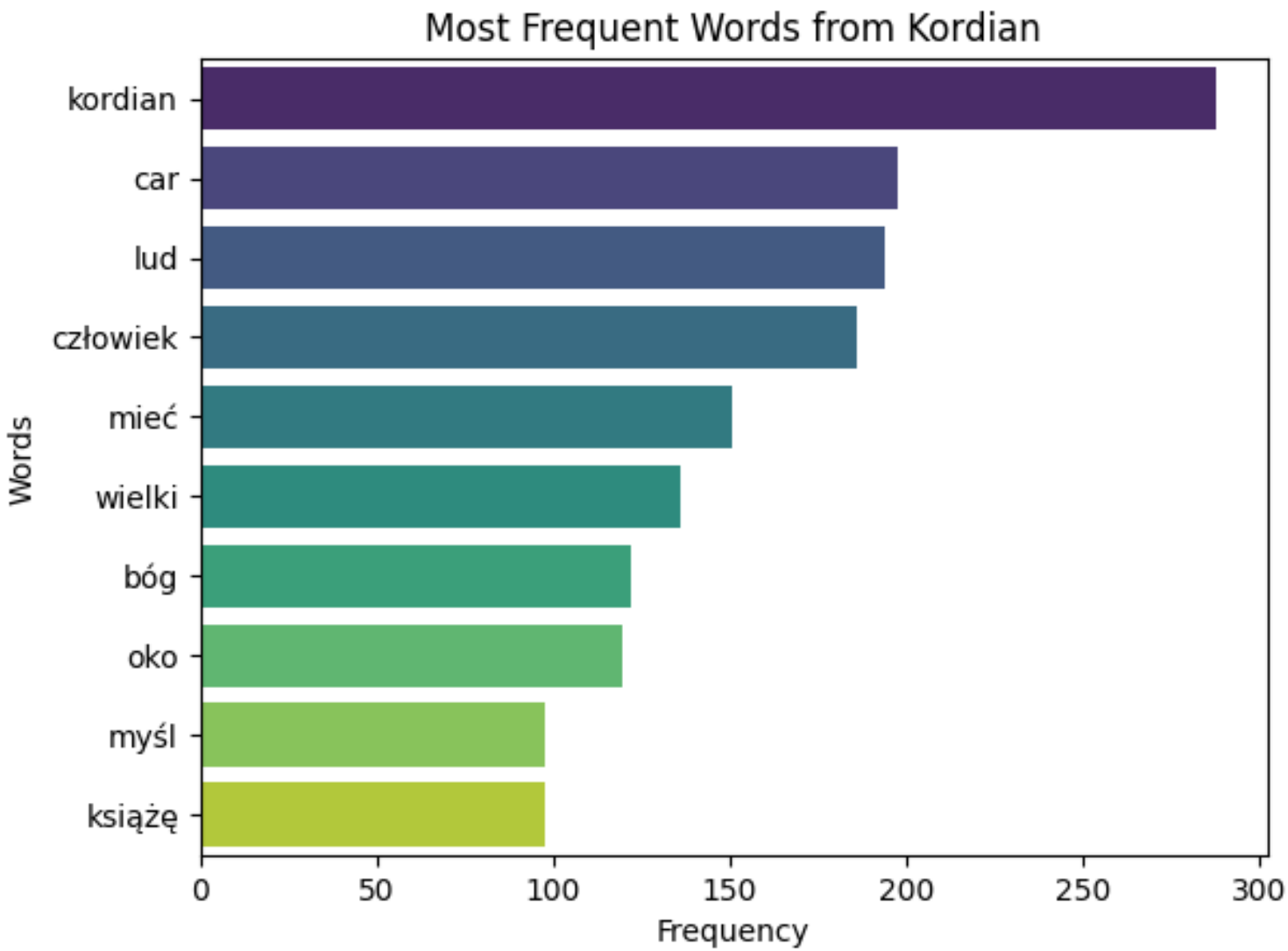
- Clear genre-based clustering
- Trilogy dominates corpus size

Word Length Distribution in Kordian



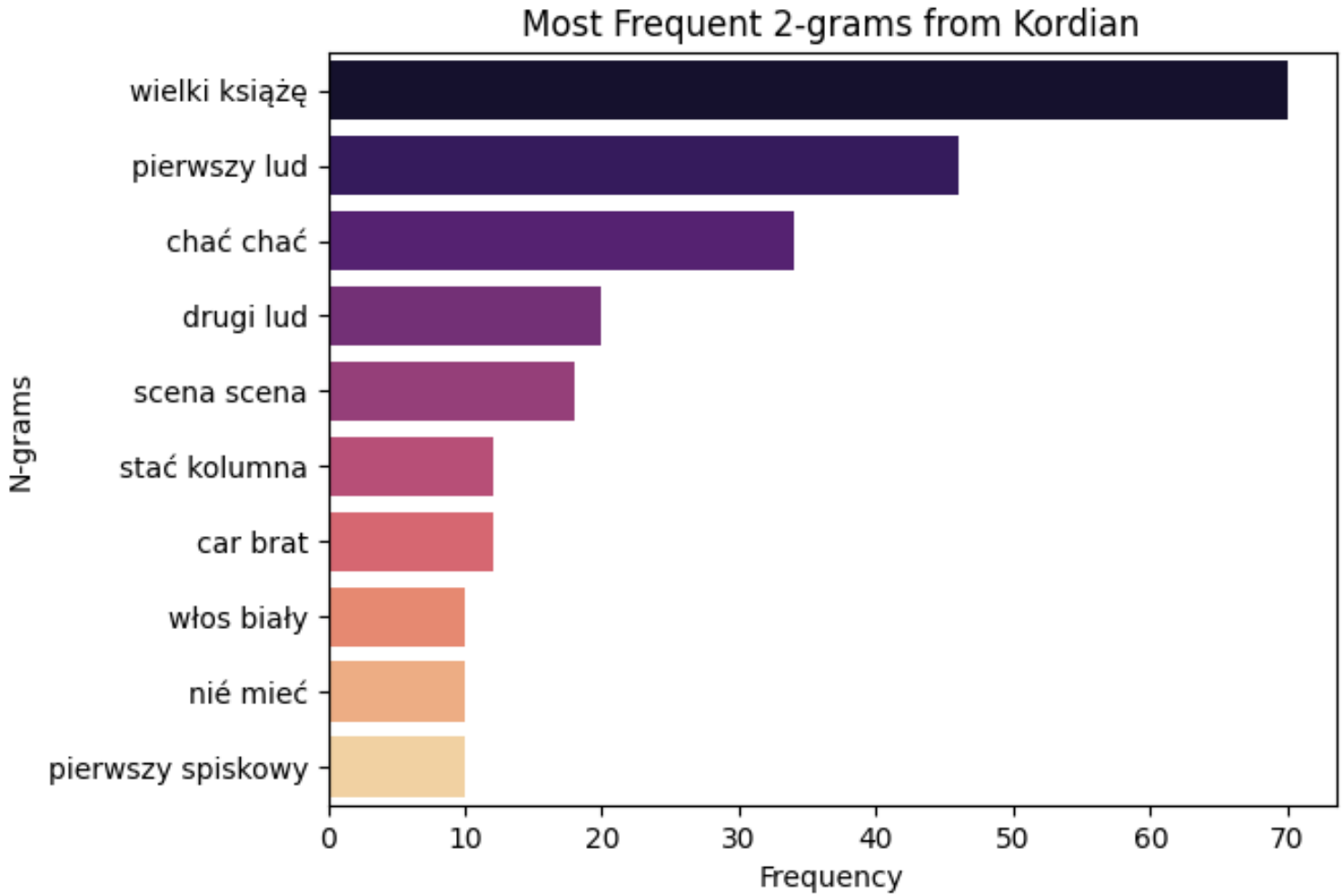
- Normal distribution pattern, peak at 5-6 charactes

Most Common Words in Kordian



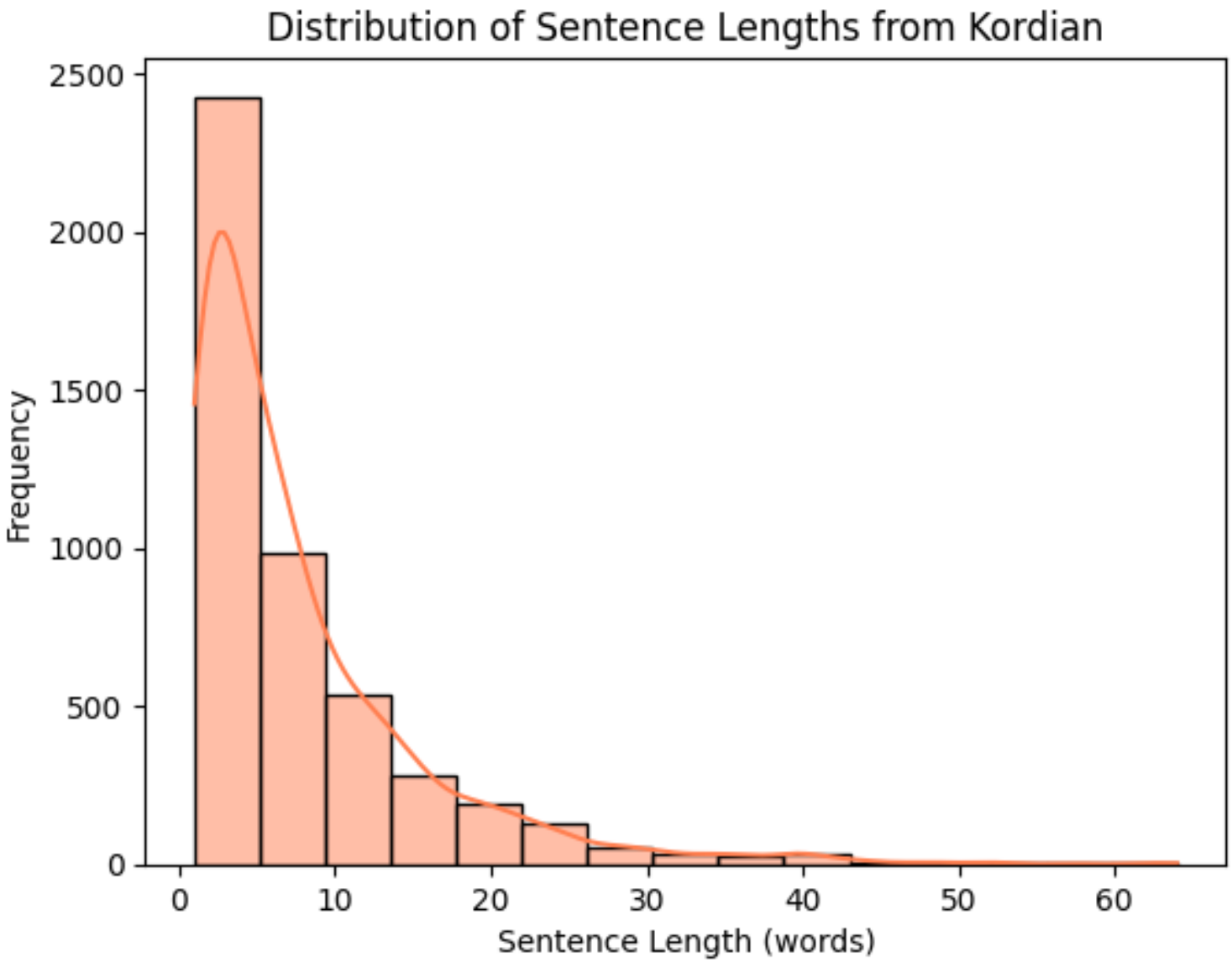
- Function words dominate
- Reflects dramatic dialogue

Most Frequent 2-grams in Kordian



- Common phrase patterns
- Stage directions present

Sentence Length Distribution in Kordian



Experiments

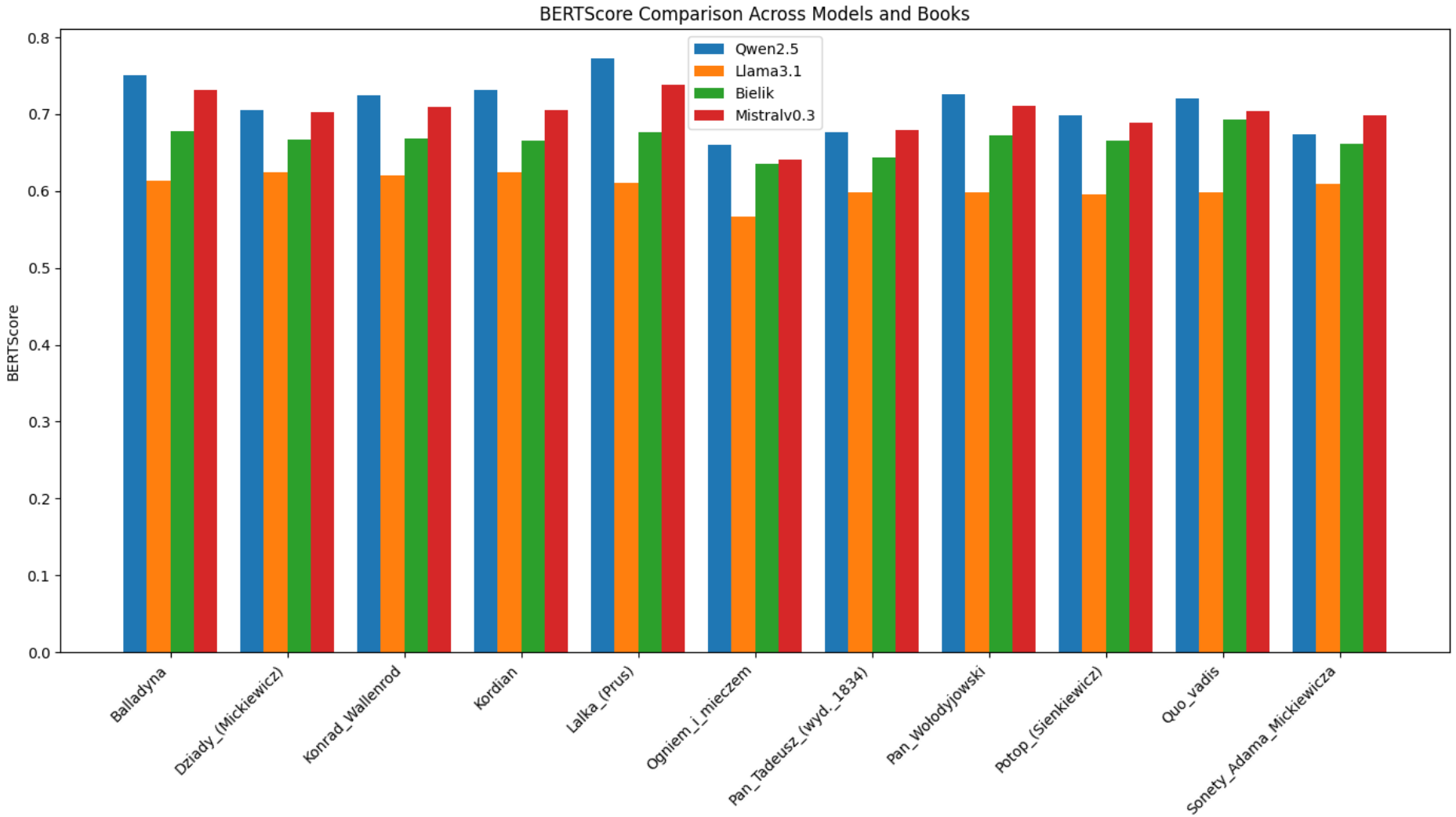
Performance Overview

- 11 datasets, all had 100 questions
- Metrics: BLEU, METEOR, ROUGE1, BERT-score
- Models: QWEN, Llama, Bielik, Mistral

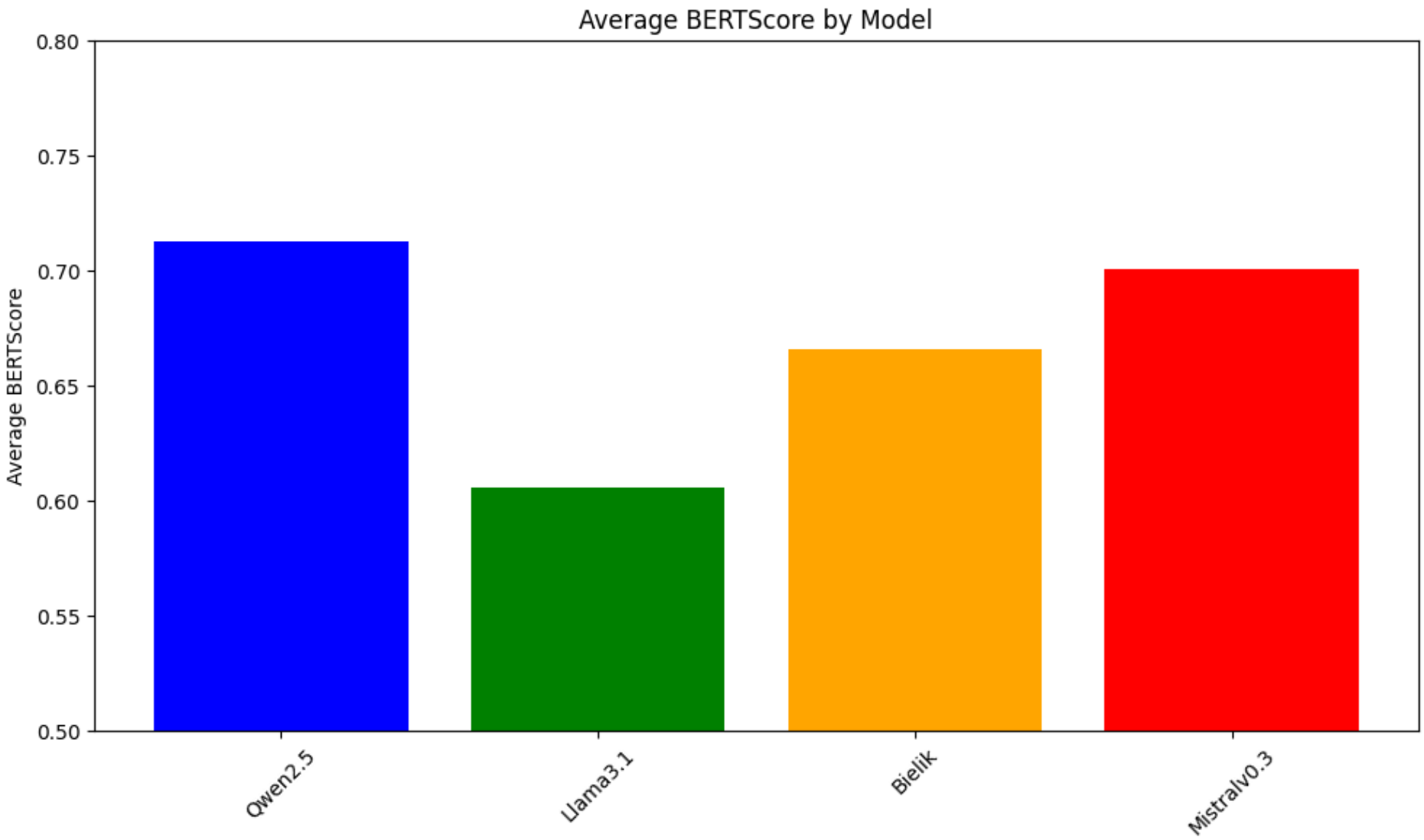
Performance Overview - QWEN

Dataset	BLEU	METEOR	ROUGE1	BERT-score
Balladyna	0.252	0.000	0.310	0.751
Dziady_(Mickiewicz)	0.155	0.000	0.221	0.705
Konrad_Wallenrod	0.158	0.000	0.234	0.724
Kordian	0.180	0.000	0.251	0.732
Lalka_(Prus)	0.305	0.000	0.368	0.772
Ogniem_i_mieczem	0.038	0.000	0.097	0.660
Pan_Tadeusz_(wyd._1834)	0.093	0.000	0.138	0.676
Pan_Wołodyjowski	0.149	0.000	0.235	0.726
Potop_(Sienkiewicz)	0.124	0.000	0.191	0.699
Quo_vadis	0.135	0.000	0.215	0.721
Sonety_Adama_Mickiewicza	0.085	0.000	0.135	0.674

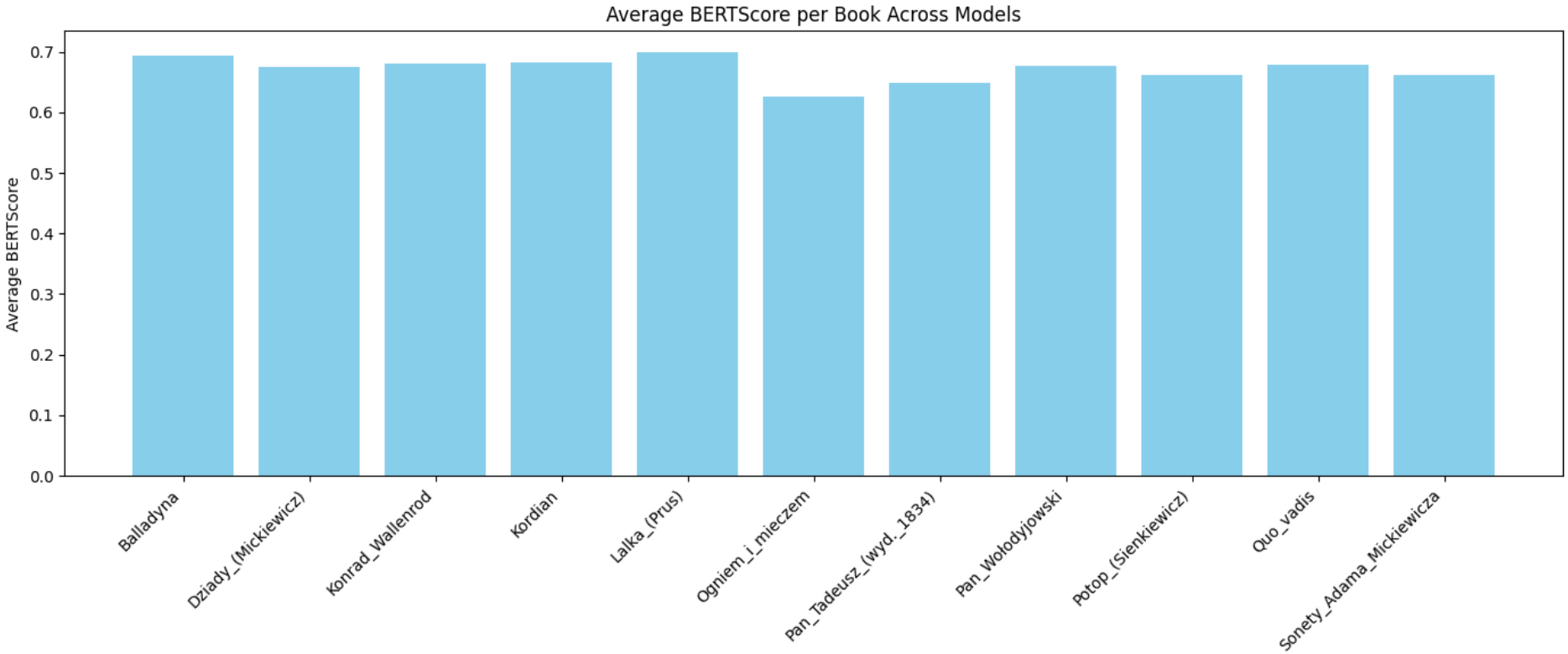
Results


BERTScore Comparison Across Models and Books
Qwen2.5
Llama3.1
Bielik
Mistralv0.3
BERTScore
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Balladyna
Dziady_(Mickiewicz)
Konrad_Wallenrod
Kordian
Lalka_(Prus)
Ogniem_i_mieczem
Pan_Tadeusz_(wyd._1834)
Pan_Wołodyjowski
Potop_(Sienkiewicz)
Quo_vadis
Sonety_Adama_Mickiewicza

Results - models



Results - books



Key Findings

- Qwen2.5 emerged as the most effective model for Polish texts, surpassing both Mistralv0.3 and the Polish-specific model Bielik, which was unexpectedly in third place.
- There is a clear need for improvement in Polish-specific LLMs and evaluation methods tailored to the language, emphasizing the importance of advanced contextual metrics.
- Future work should focus on fine-tuning Polish-specific models like Bielik with additional high-quality data and expanding datasets to address complex linguistic structures and historical vocabulary.

Key Findings

- More reference answers might increase the results, however one can do only so many references
- The results of the primary metrics are not as good as we predicted
- Traditional metrics like BLEU, METEOR, and ROUGE1 are insufficient for evaluating Polish texts due to the language's lexical diversity, highlighting the need for advanced metrics like BERTScore.
- Polish language is pretty hard to evaluate using basic metrics, more complex ones are much better

References

- Vaswani et al. (2017). "Attention Is All You Need"
- Devlin et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers"
- Ociepa et al. (2024). "Bielik 7B v0.1: A Polish Language Model"