Unveiling Narratives: Analyzing Character Roles and Themes in Novels

Jeeva Saravana Bhavanandam[1], **Subramaniya Siva T S**[1], **Om Sai Krishna Madhav Lella**[1], **Thejesh Mallidi**[1]
[1] 2nd year MS in Data Science, College of Natural Science, College of Engineering, Michigan State University

{sarava25, tsubrama, lellaom, mallidit}@msu.edu

Abstract

This project develops a natural language processing (NLP) pipeline to analyze novels, focusing on named entities, their roles, and dominant narratives. The pipeline identifies and classifies named entities as protagonists, antagonists, or other characters. It further classifies narratives within the novel using embedding-based topic modeling and generates narrative explanations through a Hybrid Retrieval-Augmented Generation (RAG) approach. A Knowledge Graph, enriched with relationships between entities and literary themes, is integrated with a VectorStore to produce coherent narrative explanations, offering deeper insights into plot progression and thematic elements within the novels.

1 Introduction

Understanding the structure of narratives and the roles of characters is a fundamental aspect of literary analysis. This project aims to develop a natural language processing (NLP) pipeline that can analyze novels by identifying key entities, such as characters, and classifying their roles within the story. In addition to identifying and categorizing these entities, the pipeline will also focus on extracting dominant narratives and providing explanations for how these narratives evolve and interact within the text.

The system will offer insights into character relationships, plot progression, and overarching themes. By focusing on both the extraction and explanation of narratives, this project seeks to shed light on the deeper connections between characters, events, and literary themes. The approach aims to provide a tool that enhances the understanding of complex narrative structures and how they shape the overall storytelling experience in novels, contributing new perspectives to the field of literary analysis.

2 Datasets

We downloaded a collection of 3,036 English books written by 142 authors from the Gutenberg Dataset and sampled 1,000 books for our analysis. We utilized the annotated data from the SemEval-2025 Task 10 to evaluate the performance of the entity role classification task.

3 Literature Survey

Recent advancements in NLP and narrative analysis have explored various aspects of entity extraction, narrative classification, and role identification, but a comprehensive system that combines these elements for novel analysis is still lacking.

3.1 Entity Role Identification

While much of the early work on narrative classification focused on analyzing persuasion and narrative framing techniques across different domains, including novels, these efforts often failed to fully integrate character role identification. Studies like (Coan et al., 2021) and (Piskorski et al., 2022) explored narrative structures but did not establish connections between entities like protagonists and antagonists with their roles in driving the narrative. In addition, works such as (Kotseva et al., 2023) advanced fine-grained narrative classification, offering methodologies relevant to literature but did not extend to linking roles with the overarching narrative framework. Our project addresses this gap by combining character role identification with narrative detection in a unified system for novels.

3.2 Narrative Theme Identification

BERTopic employs transformer embeddings to facilitate effective thematic extraction in novels, al-

lowing for a nuanced analysis of narrative structures and cultural contexts (Grootendorst, 2020). Recent studies have harnessed this approach to uncover common themes across genres and examine reader responses, thereby deepening our understanding of literary works (Khan et al., 2021; Lee Kim, 2022). In our analysis, we utilized topic modeling to identify the major themes within the narratives of the novels.

3.3 Narrative Extraction with Knowledge Graphs

Knowledge graphs have gained prominence in narrative extraction, especially for representing relationships between events and entities. (Lystopadskyi et al., 2023) developed a hybrid approach using semantic graphs to enhance narrative extraction, while (De Kok et al., 2024) introduced the concept of causality and event relations, making it highly relevant for literary plots. Additionally, (Blin, 2022) demonstrated how knowledge graphs can capture complex narrative structures in literature, which informs our approach to structuring entity relationships. Our project builds on these efforts by applying knowledge graphs to novels, mapping out the connections between characters, themes, and narrative progression.

4 Methodology

This section describes the key stages of the NLP pipeline, covering Named Entity Recognition (NER), entity role classification, narrative classification, and narrative explanation generation. Each component plays a crucial role in building a cohesive system for analyzing novels.

4.1 Named Entity Recognition (NER)

In the first step, Named Entity Recognition (NER) identifies significant entities, such as characters, locations, and objects, within the text. Transformer-based models like RoBERTa, spacy pre-trained models and custom rule-based approaches (using POS) are used to recognize and categorize these entities, forming the basis for understanding the key elements of the narrative. The accurate extraction of these entities is essential for subsequent role classification and narrative interpretation. NER performance will be evaluated using Precision, Recall, and F1-score to ensure accurate and comprehensive entity extraction.

4.2 Entity Role Classification

Once the entities are extracted, the next phase focuses on classifying their roles within the narrative. This involves determining whether an entity serves as a protagonist, antagonist, or plays a supporting role in the story. A fine-tuned transformer model analyzes the surrounding text and context of each entity to assign these roles, helping to clarify the dynamics between characters and their contributions to the plot's progression. Accuracy and Confusion Matrix will be used to evaluate the effectiveness of the entity role classification.

4.3 Narrative Theme Identification

To uncover the various themes, we split each novel into multiple smaller documents, assuming each document corresponds to a single theme. We then transformed these documents into vector embeddings using sentence-transformer models. Next, we applied UMAP and HDBSCAN to reduce the dimensionality of the embeddings and cluster the documents into common themes. We labeled each cluster as a theme by prompting the LLaMA model with the TF-IDF keywords and the most representative documents. Finally, based on the novel-to-document mapping, we identified multiple themes in each novel. The quality of the theme clusters will be assessed using the Cluster Coherence metrics.

4.4 Narrative Explanation Generation

The final stage is narrative explanation generation, achieved through a Hybrid Retrieval-Augmented Generation (RAG) approach that integrates a Knowledge Graph and a VectorStore with a large language model (LLM). The Knowledge Graph maps relationships between entities and their roles, such as interactions between protagonists, antagonists, and other characters, as well as links to key themes and plot events. This graph helps establish a structured understanding of the narrative.

The VectorStore holds semantic embeddings of the novel's sections and narrative elements, allowing for efficient retrieval of relevant content. This integration allows the system to generate coherent and contextually rich narrative explanations, combining structured data from the Knowledge Graph with unstructured text from the VectorStore. Together, these elements ensure that the generated explanations provide a deep understanding of the

novel's themes, character dynamics, and plot developments.

5 Contributions

- **Thejesh Mallidi**: Named Entity Recognition (NER) and summarization
- **Subramaniya Siva T S**: Entity Role Classification
- **Om Sai Krishna Madhav Lella**: Narrative Theme Identification
- **Jeeva Saravana Bhavanandam**: Narrative Explanation Generation

References

Coan, T. G., Ahn, W., & Lanning, J. (2021). Narrative Structure and Role Identification in Storytelling. *Journal of Narrative Theory*, 41(2), 145–172. https://doi.org/10.1353/jnt.2021.0011

Piskorski, J., Bień, J. S., & Wieczorek, K. (2022). Character Role Extraction in Literary Texts: A Comprehensive Framework. *Literary and Linguistic Computing*, 37(1), 65–87. https://doi.org/10.1093/llc/fqac011

Kotseva, M., Novak, K., & Ivanova, T. (2023). Advancing Fine-Grained Narrative Classification for Literature. *Computational Narratology Review*, 5(4), 222–234. https://doi.org/10.1093/cnr/fad001

Grootendorst, M. (2020). BERTopic: Leveraging Transformers for Topic Modeling. Available at https://github.com/MaartenGr/BERTopic

Khan, S., Weber, J., & Thakkar, M. (2021). Understanding Reader Responses through Thematic Analysis of Novels. *Digital Humanities Quarterly*, 15(3). https://doi.org/10.5555/dhq.15.3

Lee, H., & Kim, J. (2022). Cultural Contexts in Fictional Narratives: An Embedding-based Approach. *Literary Studies and Digital Methods*, 8(1), 23–38. https://doi.org/10.1007/lsdm.2022.009

Lystopadskyi, D., Santos, A., & Leal, J. P. (2023). Narrative Extraction from Semantic Graphs. In *12th Symposium on Languages, Applications and Technologies (SLATE 2023)*. OASIcs, Volume 113, pp. 9:1–9:8, Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/OASIcs.SLATE.2023.9

De Kok, M., Rebboud, Y., Lisena, P., Troncy, R., & Tiddi, I. (2024). From Nodes to Narratives: A Knowledge Graph-based Storytelling Approach. In *Seventh International Workshop on Narrative Extraction from Texts (Text2Story)*, colocated with ECIR 2024, March 24th, 2024, Glasgow, UK.

Blin, I. (2022). Building Narrative Structures from Knowledge Graphs. In Groth, P., et al. *The Semantic Web: ESWC 2022 Satellite Events*. Lecture Notes in Computer Science, vol 13384, Springer, Cham. https://doi.org/10.1007/978-3-031-11609-4_38