Introduction:

The “Goodreads” analysis project aims to identify key differences between reviews containing spoilers and those without. The team analyzed a dataset composed of general review information and spoiler-specific data from Goodreads. By employing sentiment analysis and Latent Dirichlet Allocation (LDA), the team was able to explore the influence of spoiler reviews on readers’ emotions and uncover hidden topics within the reviews. The study concluded that the spoiler reviews significantly impacted other readers’ experiences and provided insights for tailoring book themes to different audiences, enhancing online bookstores and library recommendation systems. Future work includes developing models to detect spoilers in reviews and improving sentiment analysis techniques for more nuanced insights.

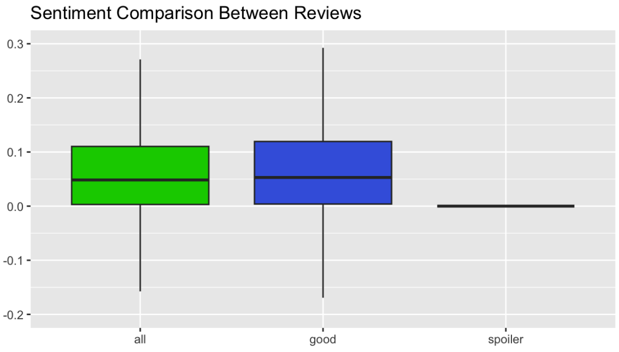
Analytical Process:

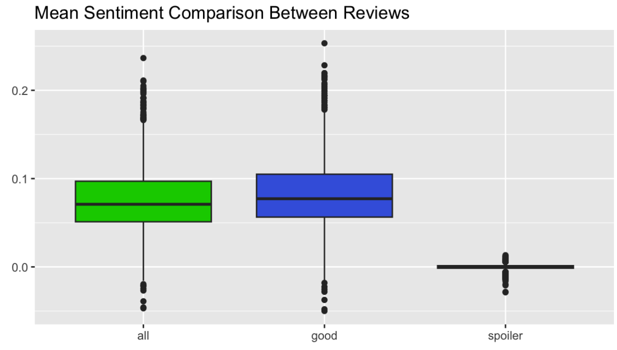
This project entails a detailed analysis of Goodreads reviews, focusing on the impact of spoilers on reader sentiment. The workflow is divided into three main phases: data preparation, sentiment analysis, and result visualization, each facilitated by a specific R script.

In the first phase, the spoiler_split.R script is used to prepare the dataset. It filters and cleans reviews for a chosen book from Goodreads. The script removes unnecessary characters from the review text, preparing it for further analysis. This step is vital for ensuring the accuracy of the sentiment analysis. From the entire dataset of over 25,000 books, this phase focuses on 2950 books to facilitate more efficient processing.

The second phase involves sentiment analysis through the sentiment.R script. Utilizing packages such as syuzhet and tidytext, this script provides a sophisticated method to gauge the emotional tone of the reviews. The “afinn” method is employed to score each word from -5 to +5, culminating in a standardized sentiment score for each review. A key feature of this script is its ability to distinguish between spoilers and non-spoiler content, offering insights into how spoilers affect reader sentiment.

In the final phase, the analyze_post.R script visualizes the analysis results. It generates boxplots and word clouds, contrasting the sentiment scores of spoiler-laden reviews with those devoid of spoilers. This visual representation highlights the neutral sentiment tendency in spoiler reviews, as opposed to the wider sentiment range in non-spoiler reviews.


Sentiment Comparison Between Reviews


Mean Sentiment Comparison Between Reviews

Utilizing the “afinn” method for sentiment analysis, we recommend books from a set of 2950, prioritizing those with high average sentiment scores and low variance, indicating consistent reader satisfaction. We also factor in the number of votes for each review, giving more weight to reviews with over 10 votes. Based on these criteria, including being in the top 20 ratings with sentiment scores above average and variance below average, our top recommendations are “Born a Crime” by Trevor Noah and “Wonder” by R.J. Palacio.

All these processes are conducted in parallel, significantly reducing the time required for completion to about three and a half hours on CHTC.

In addition to the aforementioned phases, our project integrates an advanced topic modeling component using Latent Dirichlet Allocation (LDA). This process, executed outside the main CHTC framework, enriches our analysis by uncovering underlying themes in the Goodreads reviews. LDA effectively identifies five distinct topics, ranging from 'General Book Appreciation' to 'Series and Sequels,' by analyzing word frequencies and distributions within the review text. This thematic analysis not only offers deeper insights into the content of the reviews but also enhances our understanding of how different themes may influence reader sentiment, especially in the context of spoilers. The LDA outputs, revealing predominant themes like 'Character and Story Analysis' and 'Personal and Romantic Themes,' are instrumental in dissecting the nuanced ways readers discuss and perceive various aspects of the books. Incorporating these findings into our workflow allows for a more comprehensive analysis, where sentiment scores and spoiler impacts are now contextualized within these thematic frameworks, offering a richer, multi-dimensional view of reader experiences and preferences.

Conclusion:

This method provides an in-depth analysis of sentiments in Goodreads book reviews, focusing on the effect of spoilers. This is particularly valuable for authors, publishers, and readers, as it sheds light on how spoilers in reviews impact potential readers' perceptions and interests. Such insights are crucial for understanding reader engagement and interaction in online book communities.

Future developments in literary analysis and review management show promise. These include creating specific models to detect spoilers in reviews, enhancing sentiment analysis by using phrases for more nuanced ratings, and integrating sentiment analysis with Latent Dirichlet Allocation (LDA). This integration aims to filter critical reviews, providing readers and stakeholders with more precise information about books, thereby enriching the reader experience in the digital age.

Contributions:

Members	Proposal	Coding	Presentation	Report
Philip Lin	1	1	1	1
Yanrun Lu	1	1	1	1

Xiaoyang Wang	1	1	1	1
Ziang Zeng	1	1	1	1
Chixu Ni	1	1	1	1